NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel Web Meeting

The National Quality Forum (NQF) convened a public web meeting for the Scientific Methods Panel (SMP) on February 19, 2021.

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. Chris Queram, NQF Interim President/CEO, provided welcoming remarks and appreciated the importance of the SMP's charge and commitment. Additional remarks were provided by SMP Co-Chairs, Drs. David Nerenz and Christie Teigland.

Spring 2021 Cycle Updates

Hannah Ingber, NQF Senior Analyst, reviewed the measures submitted for the upcoming Spring 2021 cycle. Of the 29 complex measures submitted for review, 18 of them were maintenance measures and 11 were new measures. Ms. Ingber also reviewed the breakdown of measures by topic area and measure type.

Discussion of Evaluation Criteria & Terminology

Dr. Ma recapped the December advisory meeting discussion. As a result of the last meeting, Accountable Care Organizations (ACOs) have been added as a care setting level in the future testing and evaluation forms. The National Provider Identifier (NPI) was not added as most SMP members were opposed to this suggestion for concerns of heterogeneity. Following the December meeting, several SMP members volunteered to further deliberate over the clarifications and changes needed for updating the NQF measure evaluation guidance.

Clarification of terminology

Data element-level test vs. performance score-level test

There is consensus among the SMP members that the current terminology of "data element-level" and "performance score-level" for testing can be confusing, especially for measure types like instrument-based and PRO-PMs. For most measure types, the SMP suggests switching from "data element-level" to "person- or encounter-level," and from "performance score-level" to "accountable entity-level" (measuring performance at the entity level such as doctor, hospital, plan, etc.).

"Data element" will still be used when referring to how the measure is constructed, including information about the value set or the direct reference code for the eCQM, along with the Quality Data Model (QDM) data type, and QDM attributes used by that data element.

There was also a discussion around validity and reliability of the data elements used to construct a measure, including claims-based measures–claims submitted with ICD-10 codes that roll up to a score. The SMP acknowledges that it is hard and adds burdens to developers to test reliability and validity of those codes. Similarly, each item of a survey usually does not meet the same reliability and validity

requirements of a multi-item scale since the items are used to construct the scale, and the latter is what is important and should be evaluated.

What is and is not a composite measure?

Current definition of composite measures in the NQF Measure Evaluation Criteria and Guidance:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures ***assessed separately for each patient*** and then aggregated into one score for an accountable entity, including all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient)

The SMP discussed that the essence of a composite measure is the presence of two or more component measures that are put together to make the composite. These component measures are measures by themselves and have their own properties of reliability and validity. A summary score such as *Consumer Reports'* 1-100 score or a multi-item scale measure (e.g., Consumer Assessment of Healthcare Providers and Systems [CAHPS]) are examples of non-composite measures because their underlying components are not designed and assessed as individual measures of an accountable entity performance. The SMP also cautioned the use of components' or domains' scores as independent measures if they are not tested and validated. Dr. Patrick Romano, SMP member, clarified that the advantage of treating an all-or-none measure as a composite is that it ensures that the all-or-none construction of the measure is conceptually and empirically sound. In addition, it allows the denominators to be constructed differently, so some patients may be eligible for three components while others are eligible for six. The composite evaluation process gives the SMP a formal mechanism for this evaluation. One member also stressed that developers should not conflate scoring methods used to calculate a composite score from the component measures with the choice of which measures to combine for an aggregate score, and why. In some instances, multiple measures are chosen to reflect a defined, pre-exiting underlying quality concept (reflective model). In other instances, the concept only exists and has a name when the measures are combined and need a label (formative model). The choice of approach used to construct the composite leads to different assumptions and tests for reliability and validity.

Finally, the SMP agreed it would be helpful for NQF to provide concrete examples of what measures are or are not composite measures and why.

Reliability criteria and minimum acceptable thresholds

Dr. Nerenz gave an overview of the various ideas the SMP members have provided with regard to clearer guidance on reliability testing and evaluation. The core of the objectives of this discussion is that measure developers and end users do not want to misclassify accountable entities based on measurement error, and unreliable measures could wrongly identify an entity as an outlier. There is a general consensus among the SMP members that the commonly used reference of 0.4 threshold (i.e., Landis & Koch that identifies a kappa statistic value of > 0.4)[1] is too low and not acceptable, and not relevant in many cases; the Landis and Koch kappa calculation only applies to the reliability of the Data Element in a survey or paper/pencil assessment instrument—NOT to the reliability of the Measure Score. Hence, the arbitrary value of "0.4" applies kappa values for data element-level (i.e., person- or encounter-level) and does not apply to an intra-class correlation coefficient (ICC), signal-to-noise ratio

[1] Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-174.

(SNR), or other scores used to demonstrate reliability of the measure score-level (i.e., accountable entity-level).

As a general principle, the SMP agreed that there are many different reliability tests, and each test could have its own rule-of-thumb guideline. Patient- or encounter-level and accountable entity-level analyses also may require different standards and thresholds.

The SMP members had a deep discussion of each method and statistics.

- Kappa statistic is often used for testing inter-rater agreement. Dr. Romano gave an example where a Kappa approach can generate surprisingly low and potentially distracting estimates when random allocation would lead to very high agreement, based on the marginal probabilities, and the two rating sources are asymmetric.
- The SMP members discussed that there are two different kinds of ICCs. The SMP members agreed that while the first approach directly tests for reliability, the second approach is important for accountable entity-level reliability if a measure is tied to payment.
    1. Comparing the variance of between-group (providers or accountable entities) random effects ("signal variance") with a variance estimate that includes within group, between-measurement effects (i.e., comparing a test period with a retest period, randomly splitting a data set into two or more data sets, or comparing two or more observers or observations of the same phenomenon).
    2. Estimating how much of the TOTAL variation in performance at the patient level is explained by those provider-level random effects.
- There was strong agreement among the SMP members that a "misclassification" or "stability of classification" (i.e., calculate rates of misclassification or reclassification) is very helpful in testing accountable entity-level reliability. A nice feature of this approach is that it requires developers to describe specific plans for classification (such as star rankings). SMP member Dr. Alex Sox-Harris mentioned that various methods and metrics for classification stability are possible.[2,3] This method might be especially important in cases where established measures of stability (split sample) are <0.70.
- Opinions among the SMP members are split regarding whether reliability should be evaluated through the lens of "intended use." In this discussion, the term "use" had multiple shades of meaning. "Use" could mean pay-for-performance vs. public reporting. "Use" could mean A specific type of analysis and structure of payment incentives or reporting groupings within either pay-for-performance (P4P) or public reporting (quintiles, deciles, "star ratings", outlier identification, etc.). "Use" could mean "accountability applications" as defined broadly by NQF vs. local quality improvement initiatives. For example, some proponents of linking reliability standards to "use" argued that measures intended for payment use should be held against a higher standard for reliability so that measurement error and random noise does not unfairly affect accountable entities, while measures for other purposes could meet a lower minimum acceptable threshold. Other SMP members disagreed and noted that some measures are sufficiently reliable to detect extreme outliers, but are not not sufficiently reliable to distinguish entities across the entire distribution of scores. Opponents of invoking "use" in assessing

[2] Staggs, VS. & Gajewski, BJ. Bayesian and frequentist approaches to assessing reliability and precision of health-care provider quality measures. *Stat Methods Med Res* 2015; 26(3).
[3] Adams, JL, et al. Physician cost profiling–reliability and risk of misclassification. *N Engl J Med* 2010; 362: 1014–1024.

reliability noted that NQF endorses measures for accountability and public reporting purposes without any restriction on specific use within those domains, and once a measure is endorsed, it is impossible to control what it is used for, so the SMP should hold all measures to the same general standards within the broad "accountability applications" domain.
- Finally, the SMP strongly supported the idea of producing a guidance table on reliability for developers, in which each method (Cronbach's Alpha, Pearson correlation, ICC, SNR, IUR, Kappa, classification stability) is listed, with the appropriateness for the level of analysis, the purpose of the testing, and an acceptable range of results.

The next tangible action for the SMP and NQF staff is to complete the guidance table on reliability and seek feedback from measure developers.

Public Comment

Caitlin Flouton, NQF Senior Analyst, opened the web meeting to allow for public comment. The comments received have been summarized below.

Comment 1: There are implementers who use the measure beyond its intended use during NQF endorsement. Sometimes a measure does not get used the way it's intended. There was an endorsement+ option at one point. The base level for use of an NQF-endorsed measure should be for accountability purposes.

Comment 2: Many measures submitted indicate exactly what they want to be used for. Sometimes a standing committee's hands are tied when reliability is demonstrated, but not for the intended use (using PIUR and IUR for example).

Comment 3: Sometimes an endorsed measure is used by private payors even if it's not demonstrated as reliable for that purpose. The commenter believes we should make sure things are reliable regardless of their use, intended or otherwise.

Next Steps

Ms. Flouton summarized the next steps for the SMP; namely, to collaborate with SMP members to capture the meeting discussions in this summary. Ms. Flouton also reminded the SMP members of important upcoming dates for the Spring 2021 cycle. Preliminary reviews and ratings of Spring 2021 measures are due by February 26, 2021. NQF staff will provide meeting materials to the SMP and measure developers the week of March 22, 2021, for the SMP Spring 2021 Measure Evaluation Meeting on March 30-31, 2021.