NATIONAL QUALITY FORUM

Scientific Methods Panel 2018 In-Person Meeting

May 16, 2018

Welcome, Introductions, and Review of Meeting Objectives

Scientific Methods Panel Members

- David Cella, PhD, (Co-Chair)
- David Nerenz, PhD (Co-Chair)
- Karen Joynt Maddox, MD, MPH (Outgoing Co-Chair)
- J. Matt Austin, PhD
- Bijan Borah, MSc, PhD
- John Bott, MBA, MSSW
- Lacy Fabian, PhD
- Marybeth Farquhar, PhD, MSN, RN
- Jeffrey Geppert, EdM, JD
- Paul Gerrard, BS, MD
- Laurent Glance, MD
- Stephen Horner, RN, BSN, MBA

Scientific Methods Panel Members (continued)

- Sherrie Kaplan, PhD, MPH
- Joseph Kunisch, PhD, RN-BC, CPHQ
- Paul Kurlansky, MD
- Zhenqiu Lin, PhD
- Jack Needleman, PhD
- Eugene Nuccio, PhD
- Jennifer Perloff, PhD
- Sam Simon, PhD
- Michael Stoto, PhD
- Christie Teigland, PhD
- Ronald Walters, MD, MBA, MHA, MS
- Susan White, PhD, RHIA, CHDA

Meeting Objectives

- Review current processes and discuss potential improvements
- Discuss conceptual definitions: Reliability and Validity
- Discuss potential changes to NQF measure evaluation criteria and guidance
- Discuss next steps for the panel

Background and Context

NQF's Scientific Methods Panel: A Stakeholder Recommendation

- Promote more consistent evaluations of the Scientific Acceptability criterion
- Reduce standing committee burden
- Hopefully—promote greater participation of consumers, patients, and purchasers on NQF standing committees

Methods Panel Charge

- Conduct **evaluation of complex measures** for the criterion of Scientific Acceptability, with a focus on reliability and validity analyses and results

- Serve in an **advisory capacity** to NQF on methodologic issues, including those related to measure testing, risk adjustment, and measurement approaches

Context for Meeting

- Terminology, methods, and even philosophy vary by discipline and expertise
- Glossary of terms: In process (part of the "toolkit")
- Threshold values: Desired by many
  - *Something we will work towards (part of the "toolkit")*
    - What are the pros/cons?
    - Thresholds for what? (e.g., statistics, approaches, measure types)
    - What information do we need?
- Today, we'll try to stay out of the weeds to the extent possible
  - **Parking Lot** *for ideas, etc.*

Context for Meeting

- No healthcare performance measure is perfect
  - *What is "good enough" for NQF endorsement?*
- NQF-endorsed measures are suitable for internal quality improvement efforts AND accountability applications
  - *At present, we do not distinguish between types of accountability applications*
- We will endeavor to come to consensus
  - *Doesn't necessarily mean unanimity*
  - *Recommendations for evaluation criteria are not binding*

Discussion of Methods Panel Processes for Measure Evaluation

Methods Panel Statistics to Date

Number of Measures	Fall 2017	Spring 2018
Evaluated by MP	8 (7 new)	21 (9 new)
Evaluated by MP co-chairs	5 (63%)	13 (62%)
Measures passed by MP	4 (50%)	8 (38%)
MP decision overturned by Standing Committee	1	*TBD*

Current Process

- A minimum of three panel members will independently evaluate each measure
  - *Assignments based on expertise, availability, need for recusal, other assigned measures*
  - *NQF provides a standard evaluation form that mirrors the rating algorithms*
- The majority recommendation from the three evaluations will serve as the overall assessment of reliability and validity

Current Process

- If there is substantial disagreement in the ratings between the three reviewers, the panel co-chairs will evaluate the measure and determine the overall recommendation
  - *Requires substantial NQF staff time*
  - *Currently, more than expected need for co-chair evaluation*
- NQF staff will compile the method's panel's ratings, evaluation, and commentary on reliability and validity and provide it to NQF's standing committees
  - *Meant to inform SC's endorsement decision*
  - *SCs can overturn the Scientific Methods Panel ratings*

Lessons Learned and Course Corrections to Date

- More information needed for evaluation
  - *For maintenance measures, staff now provides a summary of the last evaluation*
  - *Staff now provides Feasibility Scorecard (for eCQMs)*
  - *Will provide full measure specifications (fully implemented by Fall 2018)*
  - *Staff perception: submissions often do not provide enough detail about methods*
- Difficulties with the evaluation form
  - *MP members have had trouble with the form*
    - » Some revisions made between Fall and Spring cycles (revised directions; continuous numbering; reordering questions)
  - *Desire (by many) for more, not less, MP feedback provided as part of the evaluation*

Lessons Learned and Course Corrections to Date

- The evaluation process
  - *Completely independent evaluations not yet working as desired*
    - » Allow for informal discussions between evaluators (phone or e-mail), but still require separate evaluations
  - *Need for extensive review by NQF staff to ensure consistency*
    - » Incorporating phone calls as needed
  - *Additional guidance needed*
    - » For risk-adjusted measures: Inclusion (or not) of certain factors in the risk-adjustment approach should not be a reason for rejecting a measure
      - *Concerns with discrimination, calibration, or overall method of adjustment are still grounds for rejecting a measure*
    - » For all measures
      - *Incomplete or ambiguous specifications are grounds for rejecting a measure—but remember that there is an option to get clarifications, although this must be done early on*
    - » More will be coming through the "toolkit"

Still…Two Key Challenges with the Process

- Lack of consensus between panel members
  - *Excessive burden on MP members, co-chairs, and staff*
  - *Delays in workflow and confusion regarding timelines*
    - » Affects project team staff and developers
  - *Uncertainties in handoffs to Standing Committee*
- Continued dissatisfaction with the evaluation form

Addressing Lack of Consensus: Two Options to Improve Workflow

- Option #1: Keep process as is, with relatively minor changes
  - *Maintain 3 separate evaluations*
  - *Earlier resolution of issues between evaluators (e.g., go straight to calls instead of e-mailing)*
  - *Simpler process for co-chair review (e.g., calls to consider)*
- Option #2: Shift to group discussion/decision
  - *The full panel discusses all measures (in-person meeting) or subgroups of the panel discuss a subset of the measures (via webinars)*
    - » All recommendations made at the meeting
    - » Summary of the discussion is provided to the standing committee instead of providing 3-5 individual evaluations

Improving the Evaluation Form: Three Options

- Option #1: Keep the form as is, with minor changes as needed
- Option #2: Essentially, allow a "free text" evaluation
  - *Modeled after preliminary analysis done by staff prior to seating the Methods Panel*
    - » Cues about what to include
    - » "Canned" questions to consider
- Option #3: Meet somewhere in the middle
  - *Much more free text, but with some check boxes (e.g., to document how a measure does/does not meet criteria)*

Break

Reliability

Some Basic NQF Terminology

- "Healthcare performance measure" used as an umbrella term: encompasses quality measures, as well as measures of cost, resource use, and access
  - *True performance is unknown*
  - *"Performance" reflects more than just quality, access, etc. (e.g., bias, etc.)*
  - *For now, let's not worry about the label*
- "*Provider*" is another umbrella term: encompasses individual clinicians, hospitals, clinics, nursing homes, home health agencies, etc.

Some Basic NQF Terminology

- Data elements – building blocks of a measure; "variables" used to calculate a measure
  - Examples include diagnosis codes, medications, admission date, birth date, questions/items from surveys
- Measure score – the computed results of the measure
  - *Examples include rates, averages, proportions*

Current Assumptions about Reliability

- There will always be some error in performance measurement
  - *Random error affects reliability; systematic error affects validity*
- Reliability is **not a static property of a measure** (it can vary under conditions of implementation)
- Reliability is not an all-or-none property and is instead a matter of degree
  - *Considerations are scope of testing, method used, and results obtained*
- Reliability does not guarantee validity

Definitions of Reliability

- Repeatability (consistency, reproducibility, stability)
- Precision

	Repeatability	Precision
Data element	X	
Performance measure score		X

- Data Element Reliability: Repeatability and reproducibility of the data elements for the same population in the same time period
- Measure Score Reliability: Precision: Proportion of variation in the performance scores due to systematic differences across the measured entities (signal) in relation to random error (noise)

Definitions of Reliability

	Repeatability	Precision
Data element	**Current**	**Consider**
Performance measure score	**Consider**	**Current**

- New idea: importance of repeatability (stability) of the measure score
- Does it make sense to think about the precision of data elements? (or is this validity? or maybe a function of the specifications)

Questions to Consider

- Repeatability, consistency, reproducibility, stability: Are these interchangeable? Should we pick one or two?
- The idea of stability of the measure score as an important facet of reliability is new to NQF.
  - *Compared to the ability to distinguish differences, is stability as important? Less? More?*
  - *Would you expect to see <u>both</u> types of analysis for score-level testing?*
- Is it useful or helpful to use the term "signal to noise" when talking about score-level reliability?
  - *Why or why not? When?*

Questions to Consider

- Any recommendations regarding "signal-to-noise" reliability estimates in submissions?
  - *Mean and variance (or other stats such as median, percentile values, IQR, etc.)*
  - *Stratified by sample size*
- Any statement regarding signal to noise testing that is limited to providers with a minimum sample size?
- Other recommendations for submissions?
  - *Examples include: sample size calculations; when testing multiple samples, average shift in rank or proportion of units in the top or bottom quintile, along with distribution of differences; standard error of measurement*

NQF Member and Public Comment

Lunch Break

Validity

Conceptual Definition of Validity

- The correctness of measurement
  - *The extent to which one can draw correct conclusions about a particular attribute based on the results of a measure*
- The extent to which a measure assesses what it intends to measure

Current Definitions: Data Element and Measure Score Validity

- Data Element Validity
  - *Correctness of the data elements as compared to an authoritative source*
- Measure Score Validity
  - *Correctness of conclusions about quality that can be made based on the measure score (i.e., a higher score on a quality measure reflects higher quality)*

	Accuracy	Correct conclusion about performance
Data element	X	
Performance measure score		X

Current Definitions: Validity
Panel Feedback

- There is a need for some additional detail on what is meant by "what it intends to measure"
  - *Measures assess quality of care indirectly, and can vary in the degree to which measure results reflect actual underlying care quality*
  - *Need clarity and specificity from developers on the quality of care dimension the measure is intended to reflect*
- As with the overall conceptual definition of validity, there is some desire to gain insight into the extent to which a higher score on the measure actually reflects higher quality
  - *"Signal-to-noise" aspect of validity*

Other Validity Issues

- Are there any assumptions about validity that should be questions, or facets of validity that we are missing?
  - *Assumption: to be valid, a measure must be reliable*
    - » Alternative way of thinking about this is that reliability and validity are two separate and distinct characteristics of performance measures

Additional Questions to Consider

- Should we add the following ideas to our current definition?
  - *The extent to which a measure assesses what it intends to measure*
  - *Adequately distinguishes between good and poor quality*
- Does it make any sense to think about accuracy of the measure score or the correctness of conclusions about data elements?
- We ask about meaningful differences as part of assessing threats to validity. How does this relate to reliability? Is it redundant?

Break

Evaluation Criteria Discussion

Potential Changes to Evaluation Criteria and Guidance

- Should validity be considered before reliability?
  - *If so, any suggestions for how to handle specifications?*
- Minor updates to the flow of the algorithm
  - *Consider any data element testing results*
- Should NQF continue to allow developers to forego reliability testing if they demonstrate data element validity?

Next Steps: Methods "Toolkit" and Beyond

Ideas for Next Steps

- Methods "Toolkit"
  - *Definitions of important terms*
  - *Descriptions of methods for demonstrating reliability and validity*
  - *Guidance on best methods for different measure types*
  - *"Thresholds" or acceptable results (or maybe rules of thumb)*
  - *Other??*
- Article in peer-reviewed journal
  - *What? Why? Who? Where?*

Looking Ahead…

- Ideas from Parking Lot
- Smaller working groups
- Any desire for longer monthly calls?

NQF Member and Public Comment

Next Steps

Next Steps

- Monthly 1-hour calls
  - *Every 2nd Thursday of the month*
  - *Next call: June 14, 3pm ET*
- Contact information: methodspanel@qualityforum.org

Adjourn