NATIONAL QUALITY FORUM
Driving measurable health improvements together

http://www.qualityforum.org

Scientific Methods Panel Spring 2020 Meeting

April 1 – 2, 2020

Welcome, Introductions, and Disclosures of Interest

NATIONAL QUALITY FORUM

Meeting and Webinar Reminders

- Meeting breaks
- Voting Quorum
- Chat feature
- Raising hand
- Muting and unmuting your line
- If possible, do not speak on speaker phone
- Introduce yourself; we are transcribing the discussion
- Technical support

NQF Scientific Methods Panel Team

- Senior Leads
  - Ashlie Wilbon, MS, MPH, FNP-C
  - Sam Stolpe, PharmD, MPH
- Project Management
  - Mike DiVecchia, PMP
  - Hannah Ingber, MPH
  - Caitlin Flouton, MS

Scientific Methods Panel Members

J. Matt Austin, PhD	Jack Needleman, PhD
Bijan Borah, MSc, PhD	David Nerenz, PhD, Co-chair
John Bott, MBA, MSSW	Eugene Nuccio, PhD
David Cella, PhD, Co-chair	Sean O'Brien, PhD
Daniel Deutscher, PT, PhD	Jennifer Perloff, PhD
Lacy Fabian, PhD	Patrick Romano, MD, MPH
Marybeth Farquhar, PhD, MSN, RN	Sam Simon, PhD
Jeffrey Geppert, EdM, JD	Alex Sox-Harris, PhD, MS
Laurent Glance, MD	Michael Stoto, PhD
Joseph Hyder, MD	Christie Teigland, PhD
Sherrie Kaplan, PhD, MPH	Ronald Walters, MD, MBA, MHA, MS
Joseph Kunisch, PhD, RN-BC, CPHQ	Terri Warholak, PhD, RPh, CPHQ, FAPhA
Paul Kurlansky, MD	Eric Weinhandl, PhD, MS
Zhenqiu Lin, PhD	Susan White, PhD, RHIA, CHDA

Meeting Overview

NATIONAL QUALITY FORUM

Meeting Agenda

- Day 1
  - Evaluation Updates
  - Process Overview and Reminders
  - Evaluation Guidance Discussion
  - Evaluation Reminders
  - Measure Evaluations
- Day 2
  - Measure Evaluations
  - Criteria Recommendations and Evaluation Guidance
  - Next Steps

NATIONAL QUALITY FORUM

Meeting Agenda: Day 1

- Welcome, Introductions, and Disclosures of Interest
- Evaluation Updates
- Process Overview and Reminders
- Evaluation Guidance Discussion
- Evaluation Reminders
- Measure Evaluations

NATIONAL QUALITY FORUM

Meeting Materials

- Annotated agenda (provided to SMP members)
  - Identifies subgroup members, lead discussants, and those recused for specific measures
- Discussion Guide
  - Includes pertinent information from the submission
    - » Goal is to minimize need for back-and-forth with submission materials and to guide discussion so that we address critical questions/concerns
  - Measures are included in same order as the agenda
    - » By subgroup, then by rating (CNR, non-passing, passed but pulled, passed but not pulled)
  - Appendix B: Additional information provided by developers
- Background Materials
  - 2011 Testing Task Force Report
  - 2019 NQF Measure Evaluation Criteria and Guidance
  - SMP Measure Evaluation Guidance

Spring 2020 Cycle Overview

Spring 2020 Evaluation Cycle Statistics

- A total of 50 measures submitted
  - Of these, 21 were evaluated by the SMP
    - » 11 new
    - » 10 maintenance measures
- 3 subgroups of 9-10; 7 measures in each subgroup
  - 15 passed reliability AND validity
  - 2 consensus not reached (CNR) on reliability or validity
  - 4 did not pass reliability and/or validity
  - 1 pulled by NQF Staff
  - 0 pulled by SMP members (for discussion and re-vote)
- Measure Types
  - Outcome: 11
  - Intermediate clinical outcome: 2
  - Composite: 1
  - PRO-PMs: 1
  - Cost: 6

Updates

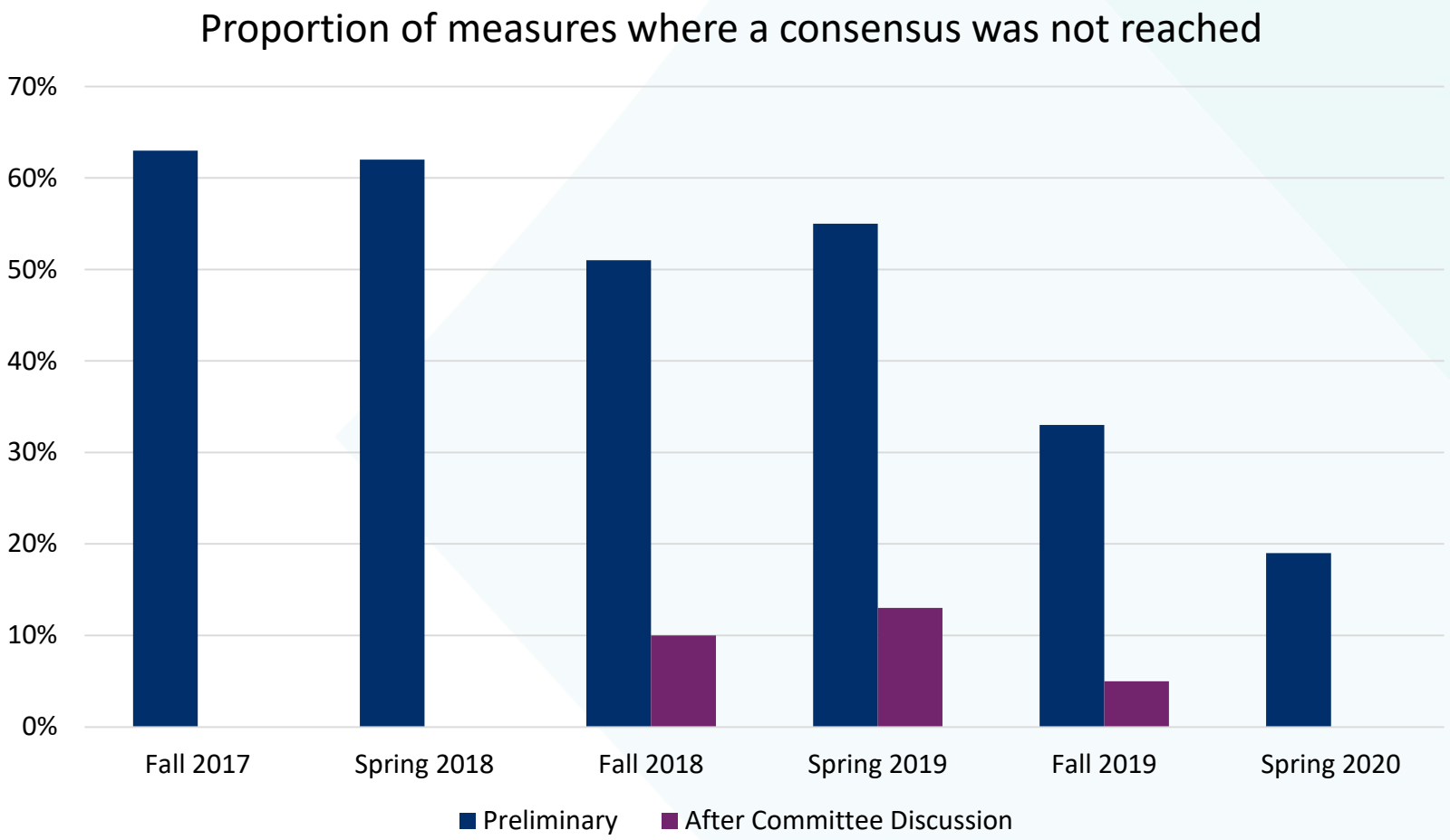
NATIONAL QUALITY FORUM

Performance Metrics

Metrics	Fall 2017	Spring 2018	Fall 2018	Spring 2019	Fall 2019	Spring 2020
Total number of complex measures submitted for evaluation by the Scientific Methods Panel (SMP)	8 (7 new)	21 (9 new)	39 (21 new)	47 (19 new)	22	21
Total Passed	4	7	25	30	17	TBD
Total Not Passed	4	13	10	11	4	TBD
Consensus Not Reached (sent to Standing committee)	0	1	4	6	1	TBD
Percent of measures where the standing committee ratings aligned with SMP recommendations (or accepted SMP ratings)	75%	100%	23/29 (79%)	35/47 (74%)	TBD	TBD

NATIONAL QUALITY FORUM

Consensus Not Reached Stats



Fall 2019 Evaluation Cycle Statistics

- 22 measures evaluated
- 15 measures discussed at meeting (68% of total)
  - 6 where consensus wasn't initially reached
  - 5 pulled by panelists for discussion
  - 4 pulled by staff for discussion
- Final results
  - Passed SMP, evaluated by SCs: n=16 (73%)
  - Consensus not reached, evaluated by SCs: n=1 (5%)
  - Did not pass: n=4 (18%)
    - » Eligible for SC re-vote: n=2 (9%)
    - » Pulled by SC for discussion: n=2 (9%)
    - » Pulled by SC for re-vote: n=2 (9%)
  - Measures withdrawn: n=1 (5%)

NATIONAL QUALITY FORUM

Panel Updates

- SMP members with Terms expiring 9/20/20
  - David Nerenz (Co-chair)
  - David Cella (Co-chair)
  - John Bott
  - Sherrie Kaplan
  - Joseph Kunisch
  - Paul Kurlansky
  - Zhenqiu Lin
  - Jennifer Perloff
  - Sam Simon
  - Michael Stoto
  - Christie Teigland

- Members can elect to continue participation for 2 years (4 cycles)
- Complete brief survey by the end of the week to indicate your decision

Meeting Overview: Process and Criteria

NATIONAL QUALITY FORUM

Complex Measure Evaluation

- All measures reviewed by the SMP can be discussed by the Standing Committees
  - Standing Committees will evaluate and make recommendations for endorsement for:
    - » Measures that pass SMP review
    - » Measures where the SMP did not reach consensus
  - Measures that did not pass the SMP can be pulled by a standing committee member for further discussion and re-vote if it is an ***eligible*** measure

Committee Consideration of Measures that Do Not Pass the SMP

- Any measure pulled by a Standing Committee member will be discussed
- Some measures may be eligible for re-vote by the Standing Committee
  - Eligibility will be determined by NQF Staff and SMP co-chairs
  - Measures that did not pass the SMP due to the following will not be eligible for re-vote:
    - » Inappropriate methodology or testing approach applied to demonstrate reliability or validity
    - » Incorrect calculations or formulas used for testing
    - » Description of testing approach, results, or data is insufficient for SMP to apply the criteria
    - » Appropriate levels of testing not provided or otherwise did not meet NQF's minimum evaluation requirements

NATIONAL QUALITY FORUM

Overall Ratings

- **High**
  - Score-level testing is required
  - A measure may be eligible for “HIGH,” but the sampling method/results may make you choose “MODERATE” instead
- **Moderate**
  - The highest eligible rating if only data element testing or face validity testing is conducted
  - A measure may be eligible for “MODERATE,” but the sampling method/results may make you choose “LOW” instead
- **Low**
  - Used primarily if testing results are not satisfactory or an inappropriate methodology was applied
- **Insufficient**
  - Use when you don’t have sufficient information to assign a “HIGH,” “MODERATE,” or “LOW” rating
    - » Example: unclear specifications; unclear testing methodology

NATIONAL QUALITY FORUM

Achieving Consensus

- Quorum: 66% of active Panel Members
- Pass/Recommended: Greater than 60% “Yes” votes of the quorum (high + moderate ratings)
- Consensus not reached (CNR): 40-60% “Yes” votes of the quorum (inclusive of 40% and 60%)
- Does not pass/Not Recommended: Less than 40% “Yes” votes of the quorum

Differences in Testing Requirements by Measure Type

- Health outcomes, intermediate clinical outcomes, cost/resource use, structure, process
  - For both reliability and validity, NQF requires **EITHER** data element testing **OR** score-level testing
    - » We prefer both, but currently do not require both
    - » Impacts rating, as described above
    - » Exception: face validity for new measures accepted
  - If data element validity testing is provided, we **do not** require additional reliability testing
    - » In this case, use the rating you give for validity as the rating for reliability
    - » This is not as common as it used to be

Differences in Testing Requirements by Measure Type

Composite measures

- NQF has specific definitions for “composite” measures
  - “Traditional” composites
  - All-or-none measures
  - Does NOT include multi-item scales in surveys/questionnaires
- Require reliability testing of the composite measure score
  - Can also show reliability testing of the components, but this is not sufficient to pass the criterion
- Score-level validity testing is not required until maintenance
- Additional subcriterion: Empirical analyses to support the composite construction
  - How this is addressed by the developer will depend on the type of composite

Differences in Testing Requirements by Measure Type

Instrument-based measures

- For reliability and validity, testing is required at **both** levels
  - Data element level → must demonstrate reliability and validity of the relevant items in the instrument, or the instrument itself
  - Measure score level → testing of the actual performance measure
- We do allow multiple performance measures under same NQF number
  - Need only one Preliminary Analysis form, but may require multiple ratings

NATIONAL QUALITY FORUM

Some Additional Reminders…

- Testing must align with specifications
  - Not a new requirement, but NQF is more rigorously upholding this requirement, particularly for level of analysis and minimum sample sizes
    - » If multiple levels of analysis are specified, each must be tested separately
  - It is possible for you to “pass” part of the measure
- Often there are several performance measures included under one NQF number
  - Each must be evaluated separately; some may pass and others may not pass

NATIONAL QUALITY FORUM

Some Additional Reminders…

- 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 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
- Empirical validity testing is expected at time of maintenance evaluation
  - If not possible, justification is required and must be accepted by the Standing Committee

NATIONAL QUALITY FORUM

Some Additional Reminders…

- Recently, the SMP articulated additional **guidance** for submissions
  1. Desire for more detail when describing methodology
  2. Desire for more than one overall statistic if reporting on signal-to-noise reliability
  3. Desire for detail in description of construct validation (e.g., narrative describing the hypothesized relationships; narrative describing why you think examining these relationships would validate the measure; expected direction of the association; expected strength of the association; specific statistical tests used; results; or interpretation of those results (including how they related to hypothesis and whether they have helped to validate the measure).
  - Lack of #2 and #3 <u>should not</u> be grounds for rejecting a measure

Methods Challenges & Measure Evaluation

NATIONAL QUALITY FORUM

Key Issues Identified during the Spring 2020 Review Cycle (and beyond)

- Reliability
  - Acceptable Thresholds
  - Acceptable methods for demonstrating reliability
  - Volume/minimum sample sizes and reliability testing
  - Relationship of reliability testing approach to validity testing approach
- Risk Adjustment
  - Social risk adjustment – further guidance on when is it appropriate?
- Cost measure evaluation challenges
  - Expectations for tailoring the risk adjustment model to the measure
  - Evaluating Exclusions

NATIONAL QUALITY FORUM

NQF Definitions of Reliability

- **Data Element Reliability**: Repeatability and reproducibility of the data elements for the same population in the same time period (consistency, reproducibility, stability)
- **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)

NATIONAL QUALITY FORUM

Reliability Discussion

- Acceptable Thresholds
  - Differing threshold values within the literature (Landis, Adams, others)
  - What is the appropriate threshold? How would the evaluation ratings be assigned based on the threshold?

Landis[1]	Adams[2]
< 0 – Less than chance agreement; 0 – 0.2 Slight agreement; 0.21 – 0.39 Fair agreement; 0.4 – 0.59 Moderate agreement; 0.6 – 0.79 Substantial agreement; 0.8 – 0.99 Almost Perfect agreement; and 1 Perfect agreement	0.5- difficult to detect differences between physicians 0.7-start to see differences between some physicians and the mean 0.9-start to see significant differences between pairs of physicians.

Examples for Reference: 0369, 1463, 3566

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

2. Estimating Reliability and Misclassification in Physician Profiling. John L. Adams, Ateev Mehrotra, Elizabeth A. McGlynn, RAND 2010

NATIONAL QUALITY FORUM

Reliability Discussion

- Acceptable methods for demonstrating reliability
  - Inter-unit (between provider) reliability (IUR) vs. Profile IUR (PIUR)
    - » How should they be interpreted when rating reliability?
    - » How should PIUR be interpreted in the context of IUR?
- Does the type of test used only demonstrate reliability for a specific purpose (e.g., detecting extreme outliers, stability)?
  - What guidance can we provide about when/how to use such methods including selection of an approach and interpretation of results?
- Volume/minimum sample sizes and reliability testing
  - Should ratings for reliability be explicitly associated with a set volume/sample size?

Morning Break (1 hour)

Reliability Discussion, Continued

- Relationship of reliability testing approach to validity testing approach
  - If the measure has been shown to only reliably be used for categorizing outliers, should the validity testing mirror this use?

NATIONAL QUALITY FORUM

Risk Adjustment Discussion

- Social risk adjustment
  - SMP can discuss the decision to include or not include social factors in the risk model but should not vote Low/Insufficient on validity solely for this reason.
    - » SMP feedback on the social risk adjustment approach is communicated to the Standing Committees, who then consider the risk model in its entirety, including clinical factors.
  - Is further guidance needed for developers and Standing Committees on determining whether factors should be included?
    - » Is the decision to include or exclude based on use and evidence surrounding what is in the provider's control?
    - » Relationship between conceptual and empirical analysis (e.g., if conceptual relationship is supported by evidence, should it be included even if minimal difference in measure scores/ranking is noted?)

NATIONAL QUALITY FORUM

Cost Measure Evaluation Challenges

- Evaluating Exclusions
  - To understand the validity of the exclusions, do we need to better understand the processes for identifying and evaluating costs to be excluded and whether these processes are systematic? Or is it sufficient for developers to justify exclusions solely through testing (which is the current requirement)?
- Risk adjustment
  - Expectations for tailoring the risk adjustment model to the measure
    - » Should we expect more tailoring of the risk adjuster to the patients and care circumstances? (Versus a standardized risk adjustment methodology)
      - *For example, diagnosis/condition leading to admission, new conditions that emerge during the period of cost or treatment being analyzed, and that could change the course or cost of treatment.*
- Examples: Measures #3564 and #3575

Opportunity for Public Comment

Afternoon Break (1 hour)

Measure Evaluation

NATIONAL QUALITY FORUM

Measure Discussion Process

- Staff will introduce the measure
- Lead discussants will summarize key concerns
- Other subgroup members are invited to comment
- Developers given 2-3 minutes for an initial response
- Discussion opened to full panel
  - Recused members cannot discuss or vote
  - Developers can respond to questions from panelists
- Final vote

The Voting Process

- Only Subgroup votes
  - Done via Poll Everywhere
  - Results from this vote will be the official vote of the SMP
- Measures not pulled for discussion: Pass with consent calendar

- NQF is considering transitioning to full Panel vote next cycle

NATIONAL QUALITY FORUM

Voting Test

#3559 Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)

- Subgroup 1
- Preliminary Voting Result:
  - Reliability: H-5; M-1; L-2; I-1 [Passed]
  - Validity: H-1; M-4; L-3; I-1 [Consensus Not Reached]
- Lead Discussants: Sherri Kaplan, Joseph Hyder
- Measure Developer: Yale/YNHH Center for Outcomes Research and Evaluation (CORE)
- Discussion Guide page 6
- For SMP discussion:
  - Could this measure be considered a composite measure?
  - Considerations for the standing Committee
    - How might this measure impact the use of related PRO-PMs?

#0715 Standardized Adverse Event Ratio for Congenital Cardiac Catheterization

- Subgroup 1
- Preliminary Voting Result:
  - Reliability: H-0, M-3, L-3, I-3 [Not Pass]
  - Validity: H-0, M-5, L-2, I-2 [Consensus Not Reached]
- Lead Discussants: Patrick Romano, Matt Austin
- Measure Developer: Boston Children's Hospital – Center of Excellence for Pediatric Quality Measurement
- Discussion Guide page 9

Opportunity for Public Comment

Adjourn

Day 2: Welcome, Review of Agenda

Scientific Methods Panel Members

J. Matt Austin, PhD	Jack Needleman, PhD
Bijan Borah, MSc, PhD	David Nerenz, PhD, Co-chair
John Bott, MBA, MSSW	Eugene Nuccio, PhD
David Cella, PhD, Co-chair	Sean O'Brien, PhD
Daniel Deutscher, PT, PhD	Jennifer Perloff, PhD
Lacy Fabian, PhD	Patrick Romano, MD, MPH
Marybeth Farquhar, PhD, MSN, RN	Sam Simon, PhD
Jeffrey Geppert, EdM, JD	Alex Sox-Harris, PhD, MS
Laurent Glance, MD	Michael Stoto, PhD
Joseph Hyder, MD	Christie Teigland, PhD
Sherrie Kaplan, PhD, MPH	Ronald Walters, MD, MBA, MHA, MS
Joseph Kunisch, PhD, RN-BC, CPHQ	Terri Warholak, PhD, RPh, CPHQ, FAPhA
Paul Kurlansky, MD	Eric Weinhandl, PhD, MS
Zhenqiu Lin, PhD	Susan White, PhD, RHIA, CHDA

NATIONAL QUALITY FORUM

Agenda for Day 2

- Welcome
- Process Review
- Measure Evaluations
- Criteria Recommendations and Evaluation Guidance
- Next Steps
- Adjourn

Measure Evaluation

NATIONAL QUALITY FORUM

Measure Discussion Process

- Staff will introduce the measure
- Lead discussants will summarize key concerns
- Other subgroup members are invited to comment
- Developers given 2-3 minutes for an initial response
- Discussion opened to full panel
  - A few people are recused: they cannot discuss or vote
  - Developers can respond to questions from panelists
- Final vote

NATIONAL QUALITY FORUM

Voting Test

#3556 National Healthcare Safety Network (NHSN) Nursing Home-onset Clostridioides difficile Infection (CDI) Outcome Measure

- Subgroup 1
- Preliminary Voting Result:
  - Reliability: H-0; M-1; L-5; I-3 [Not Pass]
  - Validity: H-0; M-1; L-5; I-3 [Not Pass]
- Lead Discussant: John Bott, Larry Glance
- Measure Developer: Centers for Disease Control and Prevention
- Discussion Guide page 10

#2496 Standardized Readmission Ratio (SRR) for Dialysis Facilities

- Subgroup 2
- Preliminary Voting Result:
  - Reliability: H-0, M-4, L-3, I-0 [Consensus Not Reached]
  - Validity: H-0, M-2, L-5, I-0 [Not Pass]
- Lead Discussants: Eugene Nuccio, Susan White
- Measure Developer: Centers for Medicare and Medicaid Services
- Discussion Guide page 12

#3566 Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities

- Subgroup 3
- Preliminary Voting Result:
  - Reliability: H-1; M-2; L-5; I-1 [Not Passed]
  - Validity: H-1; M-7; L-0; I-1 [Passed]
- Lead Discussants: Eric Weinhandl, Marybeth Farquhar
- Measure Developer: University of Michigan Kidney Epidemiology and Cost Center
- Discussion Guide page 12

Opportunity for Public Comment

Morning Break (30 minutes)

#2539 Facility 7-Day Risk-Standardized Hospital Visit Rate after Outpatient Colonoscopy

- Subgroup 3
- Preliminary Voting Result:
  - Reliability: H-4; M-3; L-1; I-0 [Passed]
  - Validity: H-1; M-3; L-3; I-1 [Consensus Not Reached]
- Lead Discussant: Alex Sox-Harris, Sean O'Brien
- Measure Developer: Centers for Medicare & Medicaid Services
- Discussion Guide page 13

#3576 Pediatric Asthma Emergency Department Use

- Subgroup 1
- Preliminary Voting Result:
  - Reliability: H-0, M-0, L-7, I-2 [Not Pass]
  - Validity: H-0, M-2, L-4, I-3 [Not Pass]
- Lead Discussants: John Bott, Daniel Deutscher
- Measure Developer: University of California, San Francisco
- Discussion Guide page 11

Evaluation Criteria Recommendations

Criterion #2: Reliability–Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) results about the quality of healthcare delivery

2a. Reliability (must-pass)

- 2a1. Precise specifications including exclusions
- 2a2. Reliability testing—data elements or measure score

NATIONAL QUALITY FORUM

NQF Definitions of Reliability

- Repeatability (consistency, reproducibility, stability)
- Precision

	Repeatability	Precision
Data element	X	no
Performance measure score	no	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)

NATIONAL QUALITY FORUM

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

NATIONAL QUALITY FORUM

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

NATIONAL QUALITY FORUM

Reliability Testing – Data Element

- Reliability of the **data elements** refers to the repeatability/ reproducibility of the data for the same population in the same time period
  - Common Approaches
    - » inter-rater/abstractor or intra-rater/abstractor studies
    - » **internal consistency for multi-item scales**
    - » **test-retest for survey items**
- Current NQF Guidance
  - All critical data elements must be tested (not just agreement of one final overall computation for all patients).
    - » At a minimum, the numerator, denominator, and exclusions (or exceptions) must be assessed and reported separately.

NATIONAL QUALITY FORUM

Current Testing Requirements

	Structure/process/outcome	Instrument-based	Composite	eCQM
Reliability	Element OR score ("short-cut"* allowed)	Element AND score	Score	Depends on how data are stored
Validity	Element OR score or face validity**	Element AND score	New: element OR score OR face validity Maintenance: Score	Element

*No reliability testing required if data element validity testing conducted and results are adequate

** Face validity allowed for new measures, but only with accepted justification at maintenance

NATIONAL QUALITY FORUM

Reliability Criteria Discussion

- Require data element AND measure score reliability testing for ALL (complex and non-complex) measures? Or EITHER with a rationale as to why they are unable to provide the other?
- Should data element validity testing continue to waive the requirement for ALL reliability testing (i.e., no reliability testing needed if data element validity testing provided)
  - Should data element reliability be required if data element validity is provided (with measure score reliability being required)?

NATIONAL QUALITY FORUM

Guidance for Reliability Testing of the Measure Score

Considering some of the common approaches for demonstrating reliability:

- Distinguishing differences and demonstrating accurate classification (signal-to-noise), between providers.
- Split half with ICC vs Split half with Pearson's or rank ordered correlations
- Split half with assessment of providers' movements across quintiles
- Bootstrapping

Panel Consideration:

- Is it necessary to perform more than one test to adequately demonstrate reliability of the measure score? Are any of these approaches sufficient to demonstrate reliability on their own?
- Are the scores from these various tests comparable (e.g., 0.7 for ICC vs. IUR)? How should they be interpreted in relationship to each other?
- What guidance can we provide about when/how to use such methods including selection of an approach and interpretation of results?

Afternoon Break (15 minutes)

Validity

Criterion #2: Validity – Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces credible (valid) results about the quality of health care delivery

2b. Validity (must-pass)

- 2b1. Validity testing—data elements or measure score
- 2b2. Justification of exclusions—relates to evidence
- 2b3. Risk adjustment—typically for outcome/cost/resource use
- 2b4. Identification of differences in performance
- 2b5. Comparability of data sources/methods
- 2b6. Missing data

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

NATIONAL QUALITY FORUM

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 (at patient level)	Correct conclusion about performance (at provider level)
Data element	X	no
Performance measure score	no	X

Validity Testing — Measure Score

- Face validity
  - Subjective determination by experts that the measure appears to reflect quality of care
    - » Empirical validity testing is expected at time of maintenance review or, justification is required.
    - » Requires systematic and transparent process by identified experts that explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good quality from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.
- Empirical testing
  - Assesses a hypothesized relationship of the measure results to some other concept; assesses the correctness of conclusions about quality

NATIONAL QUALITY FORUM

Current Guidance

Examples of validity testing of the measure score include, but are not limited to:

- testing hypotheses that the measures scores indicate quality of care (e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method);
- correlation of measure scores with another valid indicator of quality for the specific topic; or
- relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures).

Additional Context

NQF currently does not require validity testing relative to:

- An expected outcome (e.g., process measure about foot exams for patients with diabetes does not have to be correlated to a measure about foot amputation)
- Testing is not limited to other NQF-endorsed measures
- Testing does not have to use an "external" measure or dataset
  - e.g., we allow testing of an instrument-based domain measure (e.g., treated with respect") with a "global" measure from the same instrument (e.g., would you recommend this agency)
  - Recently, some concerns about "circular" testing (e.g., stability over time)

NATIONAL QUALITY FORUM

Empirical Validity Testing — Measure Score

Challenging Examples

- Comparing CAHPS measures to themselves
- Behavioral health (substance use disorder (SUD) screening versus depression and infectious disease screening) versus actual better SUD outcomes
- Cost measures comparing to other claims-based measures with the same data elements, construct vs. content validity
  - Considerations from the Cost and Efficiency Standing Committee

NATIONAL QUALITY FORUM

Panel Considerations

- Face Validity
  - Should face validity continue to be accepted as the minimum requirement for new measure submissions?
- Empirical Validity Testing
  - Do we need additional requirements regarding the "comparator"?
  - How can we enhance our guidance about score-level validation to encourage meaningful validation?
    - » Some things better than others (best practices)?
    - » Some things not really allowed?
    - » Some things maybe okay for first endorsement, but more needed for maintenance?
  - How should correlations for validity testing be interpreted? Is there an acceptable threshold or is directionality sufficient?

Tentative Next Steps for Criteria Recommendations

- May 2020: Obtain consensus recommendations from SMP during monthly call
- NQF consideration of recommendations
- Public Commenting
- ~October 2020: Present SMP recommendations to CSAC
  - CSAC may accept/reject/modify the recommendations
  - CSAC may suggest an implementation timeframe
- Winter/Early spring: Begin to publicize changes to criteria
- NOTE that NQF often allows up to a 1-year gap between changing criteria and implementing the changes
  - Likely, any SMP-recommended changes would not be required of developers until August 2021 (or even Spring Cycle 2022)

Opportunity for Public Comment

Next Steps

NATIONAL QUALITY FORUM

Process Improvement Feedback

- SMP Preliminary Analysis form
  - Review prototype for improved form
- Voting Process
- Other recommendations

NATIONAL QUALITY FORUM

White Papers

Published

- NQF Guidelines for Evaluating the Scientific Acceptability of Risk-Adjusted Clinical Outcome Measures (Larry G. et al.)
- The NQF Scientific Methods Panel (David N. et al.)

Next

- Reliability
- Social Risk Adjustment
- Assessing Validity of Cost Measures
- Others?

NATIONAL QUALITY FORUM

Next Steps

- Measure submission deadline: April 1-15
- NQF staff will summarize the relevant measure information and discussions of the SMP, and provide to the various standing committees
  - These committees will evaluate measures in the May-June timeframe
  - CSAC decisions expected in October 2020
- Next Intent to Submit deadline: August 3, 2020

NATIONAL QUALITY FORUM

2020 SMP Meetings

Meeting Date	Tentative Topic/Activity
May 26 - 1-3PM ET	Wrap up Criteria Recommendations
June 16 - 2-4PM ET	Continue reliability guidance discussion
July 21 - 2-4PM ET	Validity Testing Guidance: Choosing a comparator
August 25 - 1-3PM ET	Orientation for new members (if needed)
October 28-29 - all-day, in-person meeting	Measure Evaluation
December 8 - 1-3PM ET	TBD

NATIONAL QUALITY FORUM

Project Contact Info

- Email: MethodsPanel@qualityforum.org
- NQF phone: 202-783-1300
- Project page: http://www.qualityforum.org/Measuring_Performance/Scientific_Methods_Panel.aspx
- SharePoint site: http://share.qualityforum.org/Projects/NQF%20Scientific%20Methods%20Panel/SitePages/Home.aspx

THANK YOU.

NATIONAL QUALITY FORUM

http://www.qualityforum.org

OUR MISSION

The trusted voice driving measurable health improvements

OUR VISION

Every person experiences high value care and optimal health outcomes

OUR VALUES

Collaboration

Leadership

Passion

Excellence

Integrity