Review and Update of Guidance for Evaluating Evidence and Measure Testing

TECHNICAL REPORT

Approved by CSAC on October 8, 2013

NATIONAL QUALITY FORUM

Contents

Review and Update of Guidance for Evaluating Evidence and Measure Testing

TECHNICAL REPORT

Background

NQF endorses performance measures that are suitable for both accountability applications (e.g., public reporting, accreditation, performance-based payment, network inclusion/exclusion, etc.) as well as internal quality improvement efforts. NQF's measure evaluation criteria and subcriteria are used to determine the suitability of measures for use in these activities. Because endorsement initiates processes and infrastructure to collect data, compute performance results, report performance results, and improve and sustain performance, NQF endorsement is intended to identify those performance measures that are most likely to facilitate achievement of high quality and efficient healthcare for patients. The criteria and subcriteria also relate to the concept of "fit for purpose". For example, the clinical evidence should support use of a measure with a specific target patient population (e.g., foot care for patients with diabetes) and testing of the measure as specified indicates under what circumstances reliable and valid results may be obtained (i.e., using the measure with a specified data source and level of analysis or for the accountable entity whose performance is being measured).

Throughout the various iterations of the NQF measure evaluation criteria, the basic criteria and concepts have remained largely unchanged. However, the measure evaluation guidance—which focuses on the specificity and rigor with which the criteria are applied—has become more comprehensive and more specific over time. The guidance on measure evaluation is intended first for steering committees that evaluate performance measures and make recommendations for NQF endorsement, as well as the staff who assist them. Second, the guidance informs measure developers about how to demonstrate that a measure meets the criteria. Third, the guidance informs NQF members and the public about how measures are evaluated and informs those who use NQF-endorsed performance measures about what endorsement means.

In 2010, the NQF convened two task forces to help provide guidance for evaluating the clinical evidence and the measure testing results for reliability and validity that is submitted in support of a measure. The approved recommendations were implemented in 2011. Testing of eMeasures also was addressed in the 2011 guidance and in some subsequent draft policy statements.

Purpose

The purpose of this project was to review the implementation of the 2011 guidance on evaluating evidence and measure testing (including eMeasure testing requirements) and to propose modifications to address any major challenges. Modifications that would potentially increase consistency and clarity in the evaluation of performance measures for potential NQF endorsement also were considered. Although simplicity is desired when possible, the evaluation of evidence, reliability, and validity is complex, requiring both objective information such as the clinical evidence and testing results and steering committee.

The specific goals of the project included:

- promote consistency in evaluation across measures and projects;
- clarify common misunderstandings about the criteria and guidance;
- remain consistent with the criteria and principles from the 2011 guidance (i.e., do not change the "bar" for endorsement or the information requested for a measure submission); and
- address the current challenges with eMeasure testing.

This project was not intended to suggest changes to the basic measure evaluation criteria or to the consensus development process (CDP). Other related concerns, such as levels of endorsement, endorsement for specific applications, endorsing measures intended only for quality improvement, and definitions of multistakeholder consensus are being addressed through the Board strategic planning process, to be followed by additional work as indicated.

The Consensus Standards Approval Committee (CSAC) reviewed and discussed the measure evaluation criteria and guidance at its in-person meetings in March and July 2013, as well as in their monthly calls in May and June. A smaller subcommittee of the CSAC, formed to more thoroughly consider the issues and offer suggestions for modifications than was possible for the full CSAC, met via conference calls in June and July. The Health Information Technology Advisory Committee (HITAC) discussed eMeasure testing requirements via conference call in May 2013 and at its in-person meeting in July 2013. A subcommittee of the HITAC also was formed to offer specific recommendations regarding eMeasure testing; this subcommittee met via conference call in August 2013.

This report presents the final modifications to the 2011 guidance for evaluating evidence and measure testing (including eMeasure testing) for public and NQF member review and comment. Comments on the draft report were reviewed and the guidance was changed where indicated. This updated guidance is incorporated in a consolidated criteria and guidance document located on the NQF measure evaluation criteria web page.

Evidence

The most common issues and challenges related to implementing the 2011 guidance on evaluating the clinical evidence included:

- Measures were submitted without a summary of the quantity, quality, and consistency of the evidence from a systematic review of a body of evidence. The reasons varied across measures and developers, but the end result was that the rating scale could not be applied consistently.

Therefore, the steering committees either rated this subcriterion as insufficient evidence or relied upon their own knowledge and memory of the evidence. This resulted in inconsistency across measures and/or projects.

- Inconsistent handling of exceptions to the evidence requirement for measures that were not directly evidence-based or focused on distal process steps (e.g., document a diagnosis, order a lab test) with either indirect evidence or no empirical evidence.
- Submitted evidence was about something other than what was being measured, or provided only indirect evidence.
- A common misunderstanding was that the guidance on evidence required randomized controlled trials (RCT).

In addition, the patient-reported outcomes (PROs) project raised the question of whether NQF should apply the same evidence requirements for PROs and health outcomes.

The CSAC and its subcommittee addressed three key questions.

1. Should NQF require a systematic review of the evidence that health outcomes and PROs are influenced by healthcare processes or structures?
2. Should NQF's current guidance requiring evidence that is based on a systematic review of the body evidence to support intermediate clinical outcomes, processes, and structures be less stringent?
3. When should an exception to the evidence requirement be considered?

Health Outcomes and Patient-Reported Outcomes (PRO)

NQF has stated a hierarchical preference for performance measures of health outcomes. Current criteria require a rationale that such outcomes are influenced by healthcare processes or structures but do not require a review of the quantity, quality, and consistency of evidence. The approved recommendations from the project on PROs in Performance Measurement established that PROs should be treated the same as other health outcomes and that the CSAC should review the question of evidence requirements. PROs include health-related quality of life/functional status, symptom and symptom burden, experience with care, and health-related behaviors.

Outcomes such as improved function, survival, or relief from symptoms are the reasons patients seek care and providers deliver care; they also are of interest to purchasers and policymakers. Outcomes are integrative, reflecting the result of all care provided over a particular time period (e.g., an episode of care). Measuring performance on outcomes encourages a "systems approach" to providing and improving care. Measuring outcomes also encourages innovation in identifying ways to impact or improve outcomes that might have previously been considered not modifiable (e.g., rate of central line infection). Due to differences in severity of illness and comorbidities, not all patients are expected to have the same probability of achieving an outcome; therefore, performance measures of health outcomes and PROs are subject to the additional criterion of risk adjustment under validity.

The CSAC reaffirmed the prior guidance for health outcomes (now also applied to PROs) that requires only a rationale that the measured outcome is influenced by at least one healthcare process, service intervention, treatment, or structure.

Quantity, Quality, Consistency of the Body of Evidence and Exceptions

The CSAC also reaffirmed the criteria and guidance that calls for an assessment of the strength of the evidence from a systematic review of the body of evidence for performance measures of intermediate clinical outcomes, processes, or structures. This is consistent with the standards established by the Institute of Medicine (IOM) for systematic reviews and guidelines. The evidence should demonstrate that the intermediate outcome, process, or structure influences desired outcomes. Evidence refers to empirical studies, but is not limited to RCTs. Because endorsement sets in motion an infrastructure to address the performance measure, the intent of the evidence subcriterion is to ensure that endorsed measures focus on those aspects of care known to influence patient outcomes.

The CSAC and subcommittee also reaffirmed the need for exceptions to the evidence subcriterion. Not all healthcare is evidence-based and systematic reviews as called for by the IOM may not be currently available or the details readily accessible to obtain information on the quantity, quality, and consistency of the evidence. However, exceptions should not be considered routine and more specific guidance is needed to promote greater consistency.

Guidance for Evaluating the Clinical Evidence – Algorithm 1

Algorithm 1 presents a modified approach to guide steering committee evaluation of the evidence submitted with a performance measure. It is consistent with the prior guidance but is intended to clarify and promote greater consistency and transparency.

The key features of this proposed guidance include:

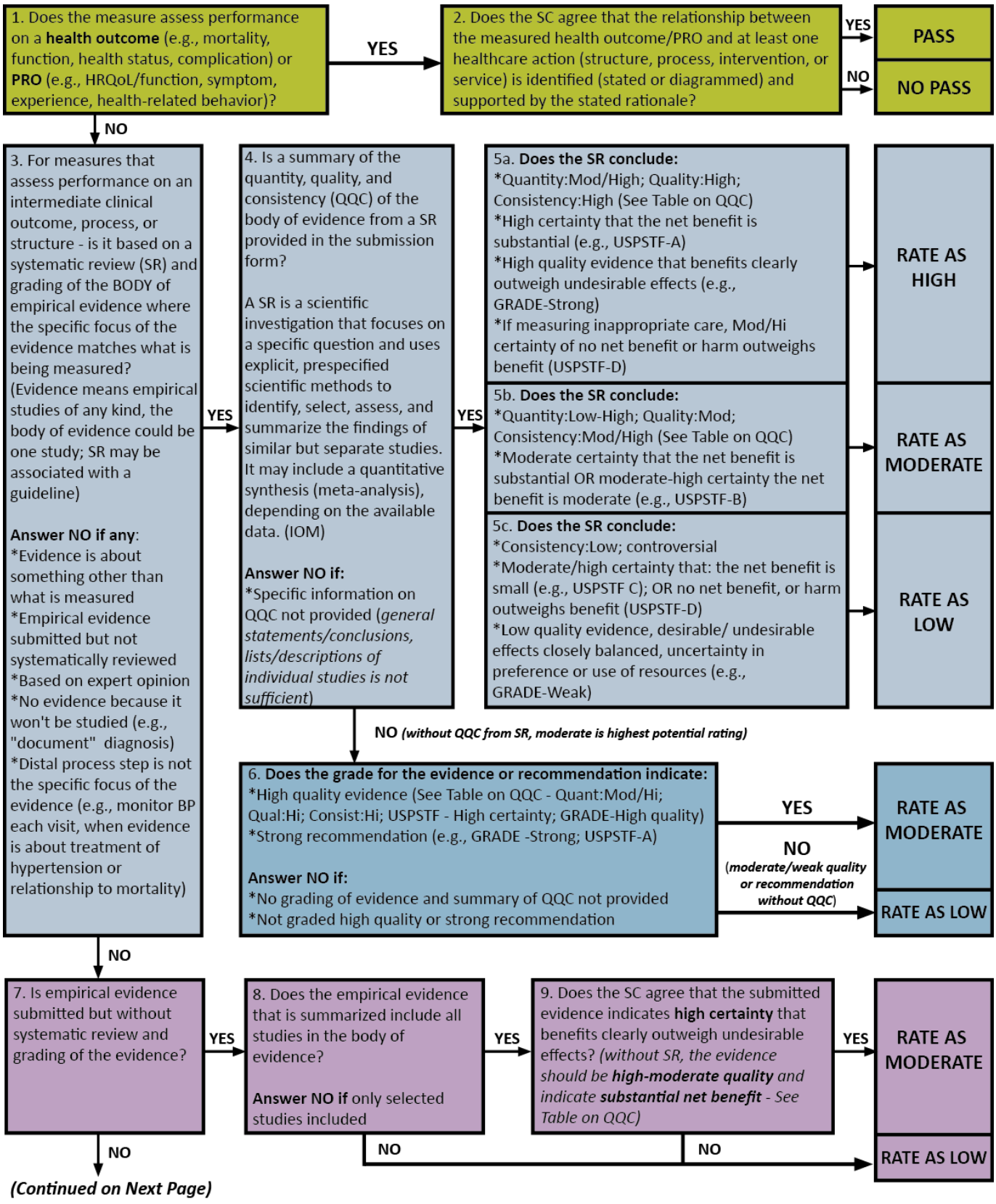
- Preserves current requirement for a rationale for measures of health outcomes and PROs.
- Preserves the basic principles of transparency and evaluating the quantity, quality, and consistency of the evidence.
- Accommodates the fact that some evidence reviews for guidelines may not be up to IOM standards or the information on quantity, quality, and consistency of the body of evidence may not be available. If evidence was graded but the submission did not include a summary of quantity, quality, and consistency, it could potentially receive a moderate rating.
- Explicitly addresses what to do if a summary of quantity, quality, and consistency of the body of evidence from a systematic review is not provided in the submission form– i.e., moderate is the highest potential rating (see boxes 4 and 6).
- Preserves flexibility for exceptions to the evidence, but identifies specific questions for considering the exception (boxes 7-9).
- Explicitly identifies how to handle measures that are based on expert opinion, indirect evidence, or distal process steps (box 3 and exceptions) and therefore need to be explicitly addressed as a potential exception.
- Uses specific examples of grades from USPSTF and GRADE in addition to the NQF rating scale (Table 1).
- The final ratings (other than for health outcomes and PROs) are high, moderate, low, and insufficient evidence and are consistent with the prior guidance where high and moderate ratings would be acceptable for endorsement. The ratings would indicate different levels of strength/certainty of the evidence, magnitude of net benefit, as well as transparency, which may be useful to implementers.
- The guidance still requires judgment of the steering committee.

The comments reflected some of the differences in perspectives about evidence requirements for measures of health outcomes and PROs. Most commenters did not address the difference in evidence requirements for outcome measures; two commenters specifically agreed with the current approach; and three commenters suggested that measures of outcomes be subject to the same evidence requirements as other measures or additional empirical analysis. In July 2013, the CSAC had reaffirmed the current criteria and guidance that requires a rationale that supports that an outcome is influenced by at least one healthcare structure, process, intervention, or service. This updated guidance was intended to reflect the current criteria, not change or "raise the bar" and the CSAC again reaffirmed the current criteria and guidance related to health outcomes and PROs as reflected in this document.

The CSAC requested comments on when exceptions to the evidence criterion should be considered. Most commenters agreed that the guidance would help promote greater consistency. In response to the comments, some revisions to the algorithm were made as follows.

- In Box 2 "plausible" rationale was replaced with a question that mirrors the criterion and the information provided in the measure submission.
- A section was added to address when there is empirical evidence that has not yet been systematically reviewed and graded (boxes 7-9).
- Some clarification was provided regarding exceptions (boxes 10-12).

Algorithm 1. Guidance for Evaluating the Clinical Evidence



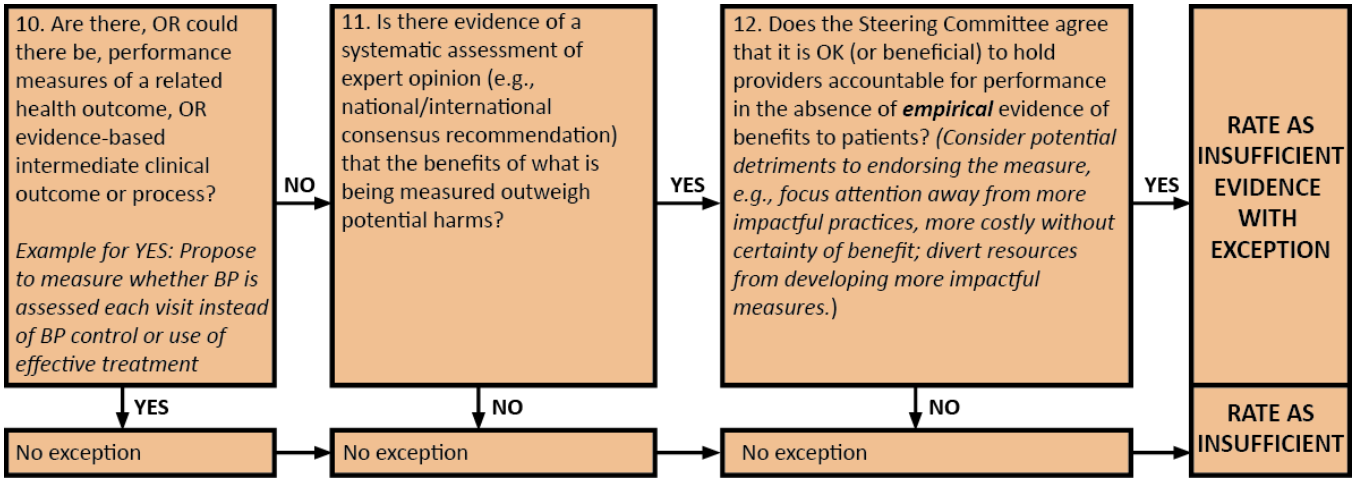

10. Are there, OR could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcome or process?
Example for YES: Propose to measure whether BP is assessed each visit instead of BP control or use of effective treatment
NO
11. Is there evidence of a systematic assessment of expert opinion (e.g., national/international consensus recommendation) that the benefits of what is being measured outweigh potential harms?
YES
12. Does the Steering Committee agree that it is OK (or beneficial) to hold providers accountable for performance in the absence of empirical evidence of benefits to patients? (Consider potential detriments to endorsing the measure, e.g., focus attention away from more impactful practices, more costly without certainty of benefit; divert resources from developing more impactful measures.)
YES
RATE AS INSUFFICIENT EVIDENCE WITH EXCEPTION
YES
No exception
NO
No exception
NO
No exception
RATE AS INSUFFICIENT

Table 1: Evaluation of Quantity, Quality, and Consistency of Body of Evidence for Structure, Process, and Intermediate Outcome Measures

DEFINITION/ RATING	**QUANTITY** OF BODY OF EVIDENCE	**QUALITY** OF BODY OF EVIDENCE	**CONSISTENCY** OF RESULTS OF BODY OF EVIDENCE
Definition	Total number of studies (not articles or papers)	Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence related to study factors[a] including: study design or flaws; directness/indirectness to the specific measure (regarding the population, intervention, comparators, outcomes); imprecision (wide confidence intervals due to few patients or events)	Stability in both the direction and magnitude of clinically/practically meaningful benefits and harms to patients (benefit over harms) across studies in the body of evidence
High	5+ studies[b]	Randomized controlled trials (RCTs) providing direct evidence for the specific measure focus, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction and similar in magnitude across the preponderance of studies in the body of evidence
Moderate	2-4 studies[b]	• Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect OR • RCTs without serious flaws that introduce bias, but with either indirect evidence or imprecise estimate of effect	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction across the preponderance of studies in the body of evidence, but may differ in magnitude If only one study, then the estimate of benefits greatly outweighs the estimate of potential harms to patients (one study cannot achieve high consistency rating)
Low	1 study[b]	• RCTs with flaws that introduce bias OR • Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations	• Estimates of clinically/practically meaningful benefits and harms to patients differ in both direction and magnitude across the preponderance of studies in the body of evidence OR • wide confidence intervals prevent estimating net benefit If only one study, then estimate of benefits do not greatly outweigh harms to patients

DEFINITION/ RATING	QUANTITY OF BODY OF EVIDENCE	QUALITY OF BODY OF EVIDENCE	CONSISTENCY OF RESULTS OF BODY OF EVIDENCE
Insufficient to Evaluate *(See Table 3 for exceptions.)*	• No empirical evidence OR • Only selected studies from a larger body of evidence	• No empirical evidence OR • Only selected studies from a larger body of evidence	No assessment of magnitude and direction of benefits and harms to patients

[a] *Study designs* that affect certainty of confidence in estimates of effect include: randomized controlled trials (RCTs), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders.

Study flaws that may bias estimates of effect include: lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; and failure to report important outcomes.

Imprecision with wide confidence intervals around estimates of effects can occur in studies involving few patients and few events.

Indirectness of evidence includes: indirect comparisons (e.g., two drugs compared to placebos rather than head-to head); and differences between the population, intervention, comparator interventions, and outcome of interest and those included in the relevant studies.[15]

[b] The suggested number of studies for rating levels of quantity is considered a general guideline.

Measure Testing

The challenges related to implementing the 2011 guidance on evaluating measure testing for reliability and validity included:

- Lack of understanding of differences between testing using patient level data versus testing using the computed performance measure score.
- Measure testing that is not consistent with the measure as specified (including data specifications and level of analysis).
- No empirical statistical testing for reliability (e.g., descriptive statistics, only report that it is in use with descriptive statistics on performance, report only a process for data management and cleaning or computer programming; report only percent agreement for inter-rater reliability).
- The rating scale did not differentiate varying levels of confidence in the results, such as when the scope of testing is narrow (e.g., 3-4 sites), or when the reliability statistic is only marginally acceptable.
- Measures were submitted for endorsement with testing results that indicated the data or the measure was not reliable or valid.
- Concerns about misclassification relate to reliability of the computed performance measure score (given that validity is demonstrated), but current criteria allow for testing of the data elements only (i.e., do not require testing at the measure score level).
- Confusion between clinical evidence for a process being measured versus validity of the performance measure as specified.
- Complexity of concepts of reliability and validity, including measure testing methods, statistical methods, and interpretation of results. Some may not be prepared to evaluate whether testing used an appropriate method, with an adequate sample, and obtained sufficient results.

- The criteria allow face validity and many measures are submitted with only face validity. Sometimes the same group of experts who helped develop the measure is used to establish face validity, or the assessment did not address the primary validity issue of whether the performance measure score from the measure as specified represents an accurate reflection of quality of care. Therefore, face validity may be questioned, especially when threats to validity such as exclusions are not adequately assessed.

The above issues also apply to eMeasures; but the most common challenges for eMeasures included:

- Measures were submitted without standard eMeasure specifications (HQMF and QDM).
- Testing that did not use electronic data (e.g., two manual abstractions).
- "Retooled" eMeasure specifications that could not be implemented.
- Difficulty recruiting test sites for testing and obtaining data from EHRs.

The CSAC and its subcommittee addressed two key questions.

- Should the rating scale reflect different levels of testing and different levels of confidence in the results?
- Can the guidance be more explicit, with recommended methods and minimum thresholds for samples and results?

In addition, the CSAC and HITAC addressed two key questions regarding eMeasures:

- Should specific thresholds for scope of testing or required type of testing be identified for eMeasures?
- How can NQF facilitate progress with eMeasures while maintaining the same criteria for endorsement as for other measures?

Testing Data Elements vs. Performance Measure Score

Data elements refer to the patient-level data used in constructing performance measures. For example, if the performance measure is the percentage of patients 65 and older with a diagnosis of diabetes with Hba1c>9 in the measurement year, then age, diagnosis (and possibly medications or lab values) are used to identify the target population of patients with diabetes for the denominator as well as potential exclusions (e.g., pregnant women) and the Hba1c lab value and date identify what is being measured for the numerator. Reliability and validity of the data elements are different from that of the computed performance score. Reliable and valid data are important building blocks for performance measures, but ultimately the computed performance measure scores are what are used to make conclusions about the quality of care provided. The question is whether the performance measure score can distinguish real differences (signal) among providers from measurement error (noise) and whether that signal is a reflection of the quality of care. These are relevant questions whether using the performance results to identify areas for improvement activities, or for purposes of accountability. The CSAC and subcommittee agreed that the rating scale should be modified slightly to reflect the difference between testing data element and performance measure scores but in such a way that the "bar" for endorsement isn't changed. For example, face validity and testing at the level of data elements should continue to be acceptable options.

More explicit Guidance on Minimum Thresholds and Types of Testing

Steering Committees often question what is considered an adequate sample for testing, and what is considered an acceptable result. However, due to the various factors and context that should be considered, the Measure Testing Task Force did not set minimum thresholds; nonetheless, they did identify some basic principles (e.g., using a representative sample of a size that was sufficient for the question and statistical method). This guidance provides much flexibility, but this flexibility can also increase uncertainty in the evaluation process and can also increase the potential for inconsistency in evaluation between measures and projects. While the CSAC and subcommittee would like to have provided some guidance regarding minimum thresholds, they again noted the difficulties in determining such thresholds and the need for steering committees to have flexibility to make judgments. For example, 0.70 is most often cited a minimum threshold for most reliability statistics, however, a higher threshold may be indicated for specific uses and 0.6 may be used for kappa.

Similarly, the Measure Testing Task Force report identified a variety of options for empirical testing and did not prescribe a particular method. The CSAC and subcommittee suggested that proposed guidance should reference the most common testing approaches but not limit measure developers from using other approaches to address the same questions.

In response to the CSAC request for comments on whether specific thresholds for the reliability statistic or sample size used in measure testing should be specified in the rating scales for reliability and validity, most commenters agreed that it is difficult or impossible to identify minimum thresholds that are applicable to all testing situations. Three commenters suggested considering power analysis to determine the appropriate sample size for the statistical test used. This will require further exploration. For now, the current guidance still applies. For eMeasures, as discussed in another section, testing should involve at least three different electronic health record systems.

Face Validity

Although empirical validity testing is preferable, NQF's criteria allow for use of face validity in lieu of empirical testing if it is systematically assessed and directed at the level of the performance measure score (i.e., whether the score is an accurate reflection of quality). Because face validity is the weakest form of validity, the proposed guidance indicated that the systematic assessment of face validity should involve experts who were not involved in measure development. Commenters requested more clarity about defining "involvement in measure development" and two commenters specifically disagreed with requiring experts beyond those involved in measure development and also noted this would "raise the bar" for meeting NQF criteria. The CSAC agreed that the proposed guidance could be considered more stringent and it would be difficult to define involvement in measure development given the variety of approaches used by different measure developers. Therefore, the updated guidance does not change requirements for face validity (i.e., developers choose who to use for systematic assessment of face validity). In the future, the CSAC will address whether face validity is sufficient for measure endorsement, especially at the time of endorsement maintenance.

Guidance on Evaluating Reliability and Validity – Algorithms 2 and 3

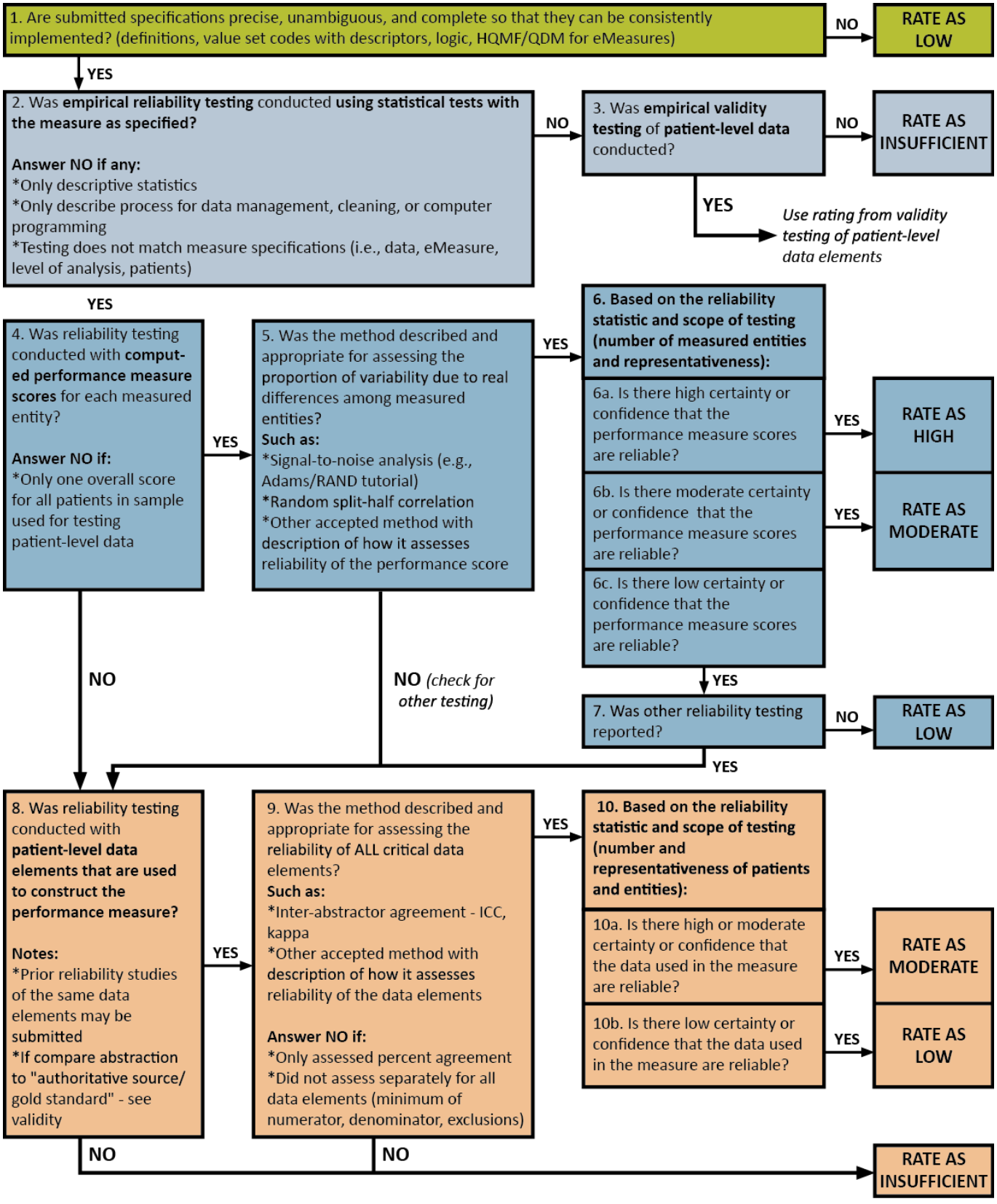
Algorithms 2 and 3 present modified approaches to guide steering committee evaluation of the reliability (Algorithm 2) and validity (algorithm 3) for all measures (including eMeasures). They are

consistent with the prior guidance but are intended to clarify and promote greater consistency and transparency.

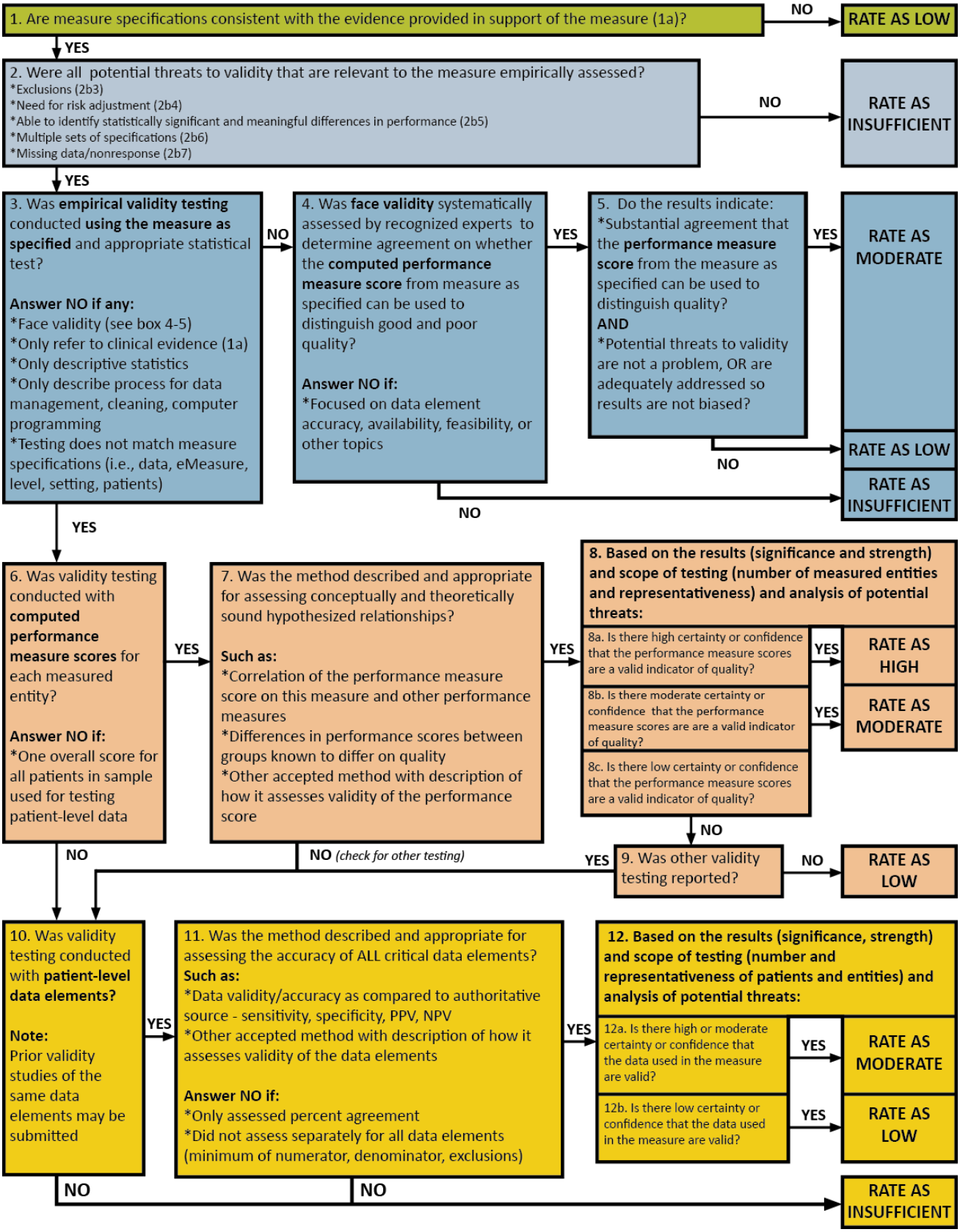
The key features of this proposed guidance include:

- Preserves most aspects of the 2011 rating scales:
  - If tested at both levels, a measure would potentially receive a high rating depending on the assessment of results and scope of testing.
  - Testing only at the level of data elements would be rated as previously – the highest potential rating is moderate, depending on results and scope of testing.
  - Face validity of the performance measure score is eligible for a moderate rating if appropriate method, scope, and result.
- The main modification to the rating scales is that testing at the level of the performance measure score alone could be eligible for a high rating, depending on result and scope of testing.
- Clarifies some common misunderstandings about testing (e.g., testing must be conducted with the measure as specified; clinical evidence is not a substitute for validity testing of the measure; data element level refers to patient-level data).
- Reinforces that testing of patient level data elements should include all critical data elements, but at minimum must include a separate assessment and results for numerator, denominator, and exclusions.
- Preserves the option to use data element validity testing for meeting both reliability and validity at the data element level.
- Reinforces that if empirical testing was not conducted or an inappropriate method was used, there is no information about reliability or validity, leading to a rating of insufficient. This preserves the distinction between insufficient information versus demonstrating low reliability or validity.
- As noted above, the general guidance on sample size for testing still applies and is considered along with appropriate method and adequate results in evaluating reliability and validity.
  - Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
  - The sample should represent the variety of entities whose performance will be measured. The Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
  - The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
  - When possible, units of measurement and patients within units should be randomly selected.

Algorithm 2. Guidance for Evaluating Reliability



Algorithm 3. Guidance for Evaluating Validity



Applying NQF Criteria for Endorsement to eMeasures

EMeasures are subject to the same evaluation criteria as other performance measures. The unique aspect of eMeasures is the measure specifications, which require the health quality measure format (HQMF) and quality data model (QDM). However, these requirements pose two significant challenges. First, the HQMF and QDM may not accommodate all types or components of performance measures (e.g., PRO-PMs, risk adjustment, composites). Second, the HQMF does not prescribe where data must be located in EHRs, usually requiring additional programming to identify where the data can be found. Therefore, it may be difficult to test eMeasures to the extent necessary to meet NQF endorsement criteria—at least until they are implemented more widely. At the same time, there is interest in developing eMeasures for use in federal programs and obtaining NQF endorsement for these eMeasures. NQF endorsement may provide the impetus to implement measures; however if a submitted measure with very limited testing does not meet NQF endorsement criteria, it could be prematurely abandoned. Some other standard-setting organizations have instituted a process to approve standards for trial use; at present, such an alternative pathway may be desirable for eMeasures.

The HITAC and CSAC requested comments on the minimum number of testing sites when conducting testing of eMeasures and whether a similar requirement should apply to all measures. Most commenters did not agree with setting a minimum for testing with three different EHRs each with three sites (total of 9 sites). They cited burden with finding sites and costs and thought it represented a "higher bar" for endorsement. Kevin Larsen clarified that ONC requires a minimum of three EHR systems, but no minimum number of sites. The CSAC and HITAC agreed to make the NQF requirement consistent with ONC and require three EHRs as reflected below.

Clarification of Requirements for Endorsing eMeasures

The following guidance addresses the criteria for endorsement of eMeasures and consolidates and clarifies the requirements for testing eMeasures submitted to NQF for endorsement (initial or endorsement maintenance). These requirements would apply to both new (de novo) eMeasures and previously endorsed measures (retooled).

- EMeasures must be specified in the accepted standard of HQMF format, and must use the Quality Data Model (QDM) and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC). Output from the Measure Authoring Tool (MAT) ensures that an eMeasure is in the HQMF format and uses the QDM (however, the MAT is not required to produce HQMF). Alternate forms of "e-specifications" other than HQMF are not considered eMeasures. *However, if HQMF or QDM does not support all aspects of a particular measure construct, those may be specified outside HQMF with an explanation and plans to request expansion of those standards.* If a value set not vetted by the VSAC explain why and plans to submit for approval. *Please contact NQF staff to discuss format for measure specifications that the standards do not support*.
- A new requirement for a feasibility assessment will be implemented with projects beginning after July 1, 2013 (see the eMeasure Feasibility Report). The feasibility assessment addresses the data elements as well as the measure logic.
- All measures (including eMeasures) are subject to meeting the same evaluation criteria that are current at the time of initial submission or endorsement maintenance (regardless of meeting prior criteria and prior endorsement status). Algorithms 1, 2, and 3 apply to eMeasures.

- Importance to Measure and Report (clinical evidence, performance gap, priority)
- Scientific Acceptability of Measure Properties (reliability, validity)
- Feasibility
- Usability and Use (Accountability/transparency, improvement)
- Related and competing measures

- To be considered for NQF endorsement, all measures (including eMeasures) must be tested for reliability and validity using the data sources that are specified. Therefore, eMeasures, whether new (de novo), previously respecified (retooled) but without eMeasure testing, or newly respecified, must be submitted with testing using the eMeasure specifications with the specified data source (e.g., EHRs, registry).
  - In the description of the sample used for testing, indicate how the eMeasure specifications were used to obtain the electronic data. Often eMeasures cannot be directly applied to EHRs or databases from EHRs and additional programming is needed to identify the location of standardized data elements. However, in some instances, the eMeasure specifications might be used directly with EHRs.
- If testing of eMeasures occurs in a small number of sites, it may be best accomplished by focusing on patient-level data element validity (comparing data used in the measure to the authoritative source). However, as with other measures, testing at the level of the performance measure score is acceptable if data can be obtained from enough measured entities. The use of EHRs and the potential access to robust clinical data provides opportunities for other approaches to testing.
  - If the testing is focused on validating the accuracy of the electronic data, analyze agreement between the electronic data obtained using the eMeasure specifications and those obtained through abstraction of the entire electronic record (not just the fields used to obtain the electronic data), using statistical analysis such as sensitivity and specificity, positive predictive value, negative predictive value. The guidance on measure testing allows this type of validity testing to also satisfy reliability of patient-level data elements (see Algorithms 2 and 3).
  - Note that testing at the level of data elements requires that all critical data elements 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.
  - Use of a simulated data set is no longer suggested for testing validity of data elements and is best suited for checking that the measure specifications and logic are working as intended.
  - NQF's guidance has some flexibility; therefore, measure developers should consult with NQF staff if they think they have another reasonable approach to testing reliability and validity.
- For eMeasures, the sample for testing the patient-level data used in constructing the eMeasures should include a **minimum of three EHR systems.**
- The general guidance on samples for testing any measure also is relevant for eMeasures:
  - Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
  - The sample should represent the variety of entities whose performance will be measured. The Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
  - The sample should include adequate numbers of units of measurement *and* adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.

- When possible, units of measurement and patients within units should be randomly selected.

- The following subcriteria under Scientific Acceptability of Measure Properties also apply to eMeasures.
  - Exclusion analysis (2b3). If exclusions (or exceptions) are not based on the clinical evidence, analyses should identify the overall frequency of occurrence of the exclusions as well as variability across the measured entities to demonstrate the need to specify exclusions.
  - Risk adjustment (2b4). Outcome and resource use measures require testing of the risk adjustment approach.
  - Differences in performance (2b5). This criterion is about using the measure as specified to distinguish differences in performance across the entities that are being measured. The performance measure scores should be computed for all accountable entities for which eMeasure data are available (not just those on which reliability/validity testing was conducted) and then analyzed to identify differences in performance.
  - eMeasures should be submitted as a separate measure even if the same or similar measure exists for another data source (e.g., claims). Therefore, comparability of performance measure scores if specified for multiple data sources (2b6) would not apply. (NQF will explore alternatives for linking measures that are the same except for data source). The measures specified for different data sources will be evaluated as competing measures (unless they apply to different care settings or levels of analysis) to determine whether one is superior to the other or whether there is justification for endorsing multiple measures.
  - Analysis of missing data (2b7). Approved recommendations from the 2012 projects on eMeasure feasibility assessment, composites, and patient-reported outcomes call for an assessment of missing data or nonresponses.

Appendix: Project Steering Committee and NQF Staff

Consensus Standards Approval Committee Member Roster

Frank Opelka, MD, FACS (Chair) *
Vice President for Health Affairs and Medical Education
Louisiana State University, New Orleans, LA

Cristie Upshaw Travis (Vice-Chair) *
Chief Executive Officer
Memphis Business Group on Health, Memphis, TN

Andrew Baskin, MD *
National Medical Director for Quality and Provider Performance Measurement
Aetna, Blue Bell, PA

Pamela Cipriano, PhD, RN NEA-BC, FAAN
Senior Director
Galloway Consulting, Marietta, GA

William Conway, MD
Senior Vice President and Chief Quality Officer
Henry Ford Health System, Detroit, MI

Robert Ellis
Director of Operations and Online Services
Consumers' Checkbook, Ashburn, VA

Lee Fleisher, MD *
Robert D. Dripps Professor and Chair of Anesthesiology and Critical Care
University of Pennsylvania, Philadelphia, PA

David Knowlton, MA
President and Chief Executive Officer
The New Jersey Health Care Quality Institute, Pennington, NJ

Philip E. Mehler, MD
Chief Medical Officer and Director of Quality
Denver Health, Denver, CO

Ann Monroe
President
Health Foundation for Western & Central New York, Buffalo, NY

Arden Morris, MD, MPH, FACS
Associate Professor of Surgery
University of Michigan Health System, Ann Arbor, MI

Lyn Paget, MPH
Managing Partner
Health Policy Partners, Boston, MA

Carolyn Pare
President and Chief Executive Officer
Buyers Health Care Action Group, Bloomington, MI

Lee Partridge *
Senior Health Policy Advisor
National Partnership for Women & Families, Washington, DC

Kyu Rhee, MD, MPP
Vice President of Integrated Health Services
IBM Corporation, Somers, NY

David Rhew, MD
Chief Medical Officer and VP of Global Healthcare
Samsung SDS America, Moonachie, NJ

Dana Gelb Safran, ScD *
Senior Vice President for Performance Measurement and Improvement
Blue Cross Blue Shield of Massachusetts, Boston, MA

David Shahian *
Chair of 2010 Evidence Task Force
Consultant Surgeon
Massachusetts General Hospital (MGH), Boston, MA

*** Participated on subcommittee**

Health Information Technology Advisory Committee Roster

Paul C. Tang, MD, MS (Chair) *
Vice President and Chief Medical Information Officer
Palo Alto Medical Foundation, Palo Alto, CA

J. Marc Overhage, MD, PhD (Vice-Chair) **
Chief Medical Informatics Officer
Siemens Healthcare, USA, Malvern, PA

Kristine Martin Anderson, MBA **
Senior Vice President
Booz Allen Hamilton, Rockville, MD

David W. Bates, MD, MSc
Medical Director of Clinical and Quality Analysis
Partners Healthcare System, Inc., Boston, MA

Zahid Butt, MD, FACG *
President and CEO, Medisolv, Inc.
Columbia, MD

Ian Z. Chuang, MD, MS *
Senior Vice President, Healthcare Informatics, and Chief Medical Officer
Netsmart, Overland Park, KS

John Derr, RPh
Health Information Strategy Consultant, Golden Living, LLC
Anacortes, WA

Richard Dutton, MD, MBA *
Executive Director, Anesthesia Quality Institute
Park Ridge, IL

Jamie Ferguson *
Vice President, Health Information Technology Strategy & Policy
Kaiser Permanente, Oakland, CA

Paul Fu, MD, MPH
Chief Medical Information Officer, Harbor - UCLA Medical Center
Torrance, CA

Leslie Kelly Hall
Senior Vice President
Healthwise, Inc., Boise, Idaho

Allison Jackson, MS
Project Manager/Epidemiologist
Intel Corporation, Chandler, AZ

Caterina E.M. Lasome, PhD, MSN, MBA, MHA, RN, CPHIMS
President and CEO
iON Informatics, LLC, Dunn Loring, VA

Russell Leftwich, MD *
Chief Medical Informatics Officer, Office of eHealth Initiatives
State of Tennessee, Nashville, TN

Michael Lieberman, MD **
Associate Chief Health Information Officer
Oregon Health Science University, Portland, OR

Andrew Litt, MD
Chief Medical Officer
Dell Healthcare and Life Sciences, Park City, UT

Erik Pupo, CPHIMS
Senior Manager, Deloitte Consulting LLP
Health Sciences and Government, Alexandria, VA

Christopher Queram, MA
President and CEO
Wisconsin Collaborative for Healthcare Quality, Madison, WI

Carol Raphael, MPA, M.Ed.
Advanced Leadership Fellow at Harvard; Former President and Chief Executive Officer
Visiting Nurse Service of New York (VNSNY), New York, NY

Deborah A. Reid, JD, MHA
Senior Attorney
National Health Law Program, Washington, DC

Joyce Sensmeier, MS, RN-BC, CPHIMS, FHIMSS, FAAN
Vice President, Informatics
Healthcare Information and Management Systems Society (HIMSS), San Diego, CA

Shannon Sims, MD, PhD
Director of Clinical Informatics
Rush University Medical Center, Chicago, IL

Christopher Snyder, DO
Chief Medical Information Officer/Chief Quality Officer
Peninsula Regional Medical Center, Salisbury, MD

Christopher Tonozzi, MD *
Chief Medical Information Officer
Colorado Associated Community Health Information Enterprise, Denver, CO

Madhavi Vemireddy, MD
Chief Medical Office and Head of Product Management
ActiveHealth Management, New York, NY

Judith Warren, PhD, RN, BC, FAAN, FACMI *
Retired, Professor
University of Kansas School of Nursing, Kansas City, KS

Federal Liaisons

Joseph Francis, MD, MPH
Director of Health Performance Measurement, Office of Informatics and Analytics
Veterans Health Administration, Washington, DC

Erin Grace, MHA
Senior Manager, Health IT
Agency for Healthcare Research and Quality, Rockville, MD

Christopher Lamer, PharmD
Medical Informaticist, Office of Information Technology
Indian Health Service, Rockville, MD

Kevin Larsen, MD
Medical Director Meaningful Use
Office of the National Coordinator for Health IT, Washington, DC

Martin Rice, MS, RN-BC
Deputy Director, Office of HIT and Quality
Health Resources and Services Administration, Rockville, MD

* Participated on subcommittee
** Also participated on CSAC subcommittee

NQF Staff

Helen Burstin, MD, MPH
Senior Vice President
Performance Measurement

Karen Beckman Pace, PhD, RN
Senior Director
Performance Measurement

Christopher Millet, MS
Senior Director
Performance Measurement

Karen Johnson, MS
Senior Director
Performance Measurement

Reva Winkler, MD, MPH
Senior Director
Performance Measurement

Taroon Amin, MA, MPH
Senior Director
Performance Measurement

Evan Williamson, MPH, MS
Project Manager
Performance Measurement

Jessica Weber, MPH
Project Manager
Performance Measurement