NATIONAL QUALITY FORUM

Memo

TO: NQF Members and Public

FR: NQF Staff

RE: Review and Update of Guidance for Evaluating Evidence and Measure Testing (including eMeasures)

DA: August 8, 2013

Background

In 2010, the NQF convened two task forces to help provide guidance for evaluating the clinical evidence and measure testing for reliability and validity that is submitted in support of a performance measure being considered for endorsement. The approved recommendations were implemented in 2011. Testing of eMeasures also was addressed in the 2011 guidance and in some subsequent draft policy statements. Some challenges and inconsistencies in applying the guidance have been identified.

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.

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.

The consensus Standards Approval Committee (CSAC), Health IT Advisory Committee (HITAC) and subcommittees of both groups worked with NQF staff from March to August 2013 to consider the issues and propose potential solutions.

Review and Comment

The CSAC and HITAC recommendations are included in the draft document *Review and Update of Guidance for Evaluating Evidence and Measure Testing*. The draft report is posted on the NQF web site for purposes of review and comment only and is not intended to be used for voting purposes.

The report presents some potential modifications to the 2011 guidance for evaluating evidence and measure testing (including eMeasure testing) for public and NQF member review and comment. A proposal for potentially approving eMeasures for trial use also is presented.

The proposed modifications basically represent the prior guidance on rating the evidence, reliability, and validity in stepwise algorithms along with some specific clarifications. They do not change the criteria or the information requested on submission, and they are not intended to change the “bar” for endorsement. The evidence algorithm provides explicit guidance on how to consider potential exceptions to the evidence and what to do when a summary of the quantity, quality, and consistency of the body evidence from a systematic review is not provided. The main modification to the rating scales for reliability and validity is that testing at the level of the performance score alone could be eligible for a high rating, depending on the results and scope of testing.

The report specifically identifies how the criteria apply to eMeasures and replaces all prior statements or guidance about eMeasure testing. A requirement for a minimum number of sites for testing data element validity of eMeasures is proposed. Additionally, because of the challenges associated with implementing eMeasures, recruiting test sites, and obtaining sufficient amounts of data, an alternative pathway with approval for trial use is proposed. This would NOT be considered endorsement, it is NOT time-limited endorsement, and is NOT a required first stage.

The CSAC and HITAC are interested in your comments before finalizing the guidance that will be provided to steering committees.

You may post your comments and view the comments of others on the NQF website using the online submission process.

All comments must be submitted no later than 6:00 PM ET, August 30, 2013.

Thank you for your interest in the NQF's work. We look forward to your review and comments.

Review and Update of Guidance for Evaluating Evidence and Measure Testing

DRAFT TECHNICAL REPORT FOR REVIEW

August 8, 2013

NATIONAL QUALITY FORUM

Contents

Review and Update of Guidance for Evaluating Evidence and Measure Testing

DRAFT 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.

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 some potential modifications to the 2011 guidance for evaluating evidence and measure testing (including eMeasure testing) for public and NQF member review and comment. Also included in this report is a proposal for another pathway to endorsement for eMeasures. The associated criteria and prior guidance are provided in the appendices.

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 judgment to review and reach a conclusion regarding what is sufficient to recommend a performance measure for NQF endorsement.

Evidence

The most common issues and challenges related to implementing the 2011 guidance on evaluating the clinical evidence (Appendix A) 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 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.

Proposed Guidance for Evaluating the Clinical Evidence – Algorithm 1

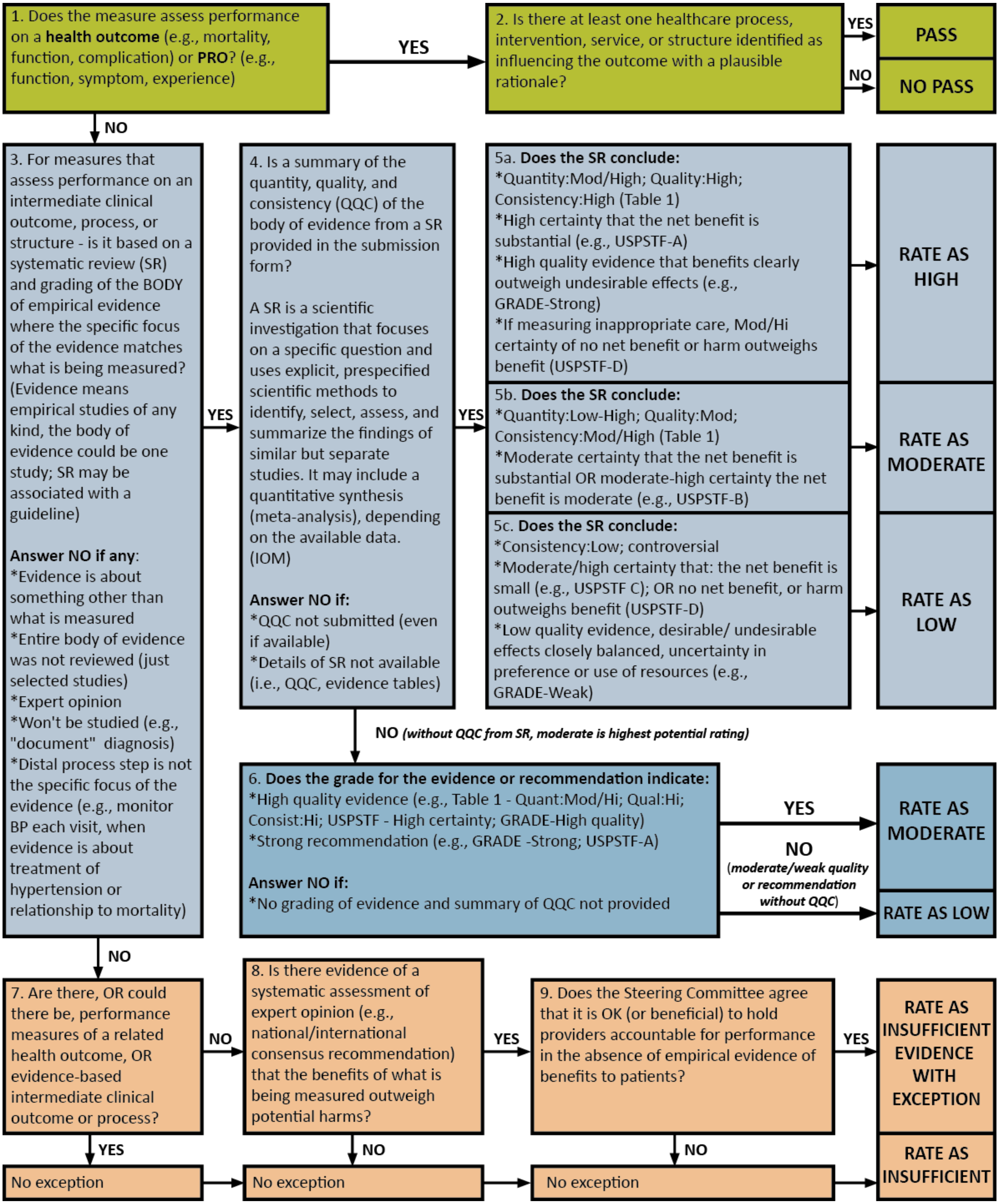
Algorithm 1 presents a modified approach to guide steering committee evaluation of the evidence submitted with a performance measure. It is consistent with the prior guidance (Appendix A) but is intended to clarify and promote greater consistency and transparency.

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 CSAC is particularly interested in receiving comments on when exceptions to the evidence criterion should be considered.

Algorithm 1. Guidance for Evaluating the Clinical Evidence



1. Does the measure assess performance on a **health outcome** (e.g., mortality, function, complication) or **PRO**? (e.g., function, symptom, experience)

2. Is there at least one healthcare process, intervention, service, or structure identified as influencing the outcome with a plausible rationale?

3. For measures that assess performance on an intermediate clinical outcome, process, or structure - is it based on a systematic review (SR) and grading of the BODY of empirical evidence where the specific focus of the evidence matches what is being measured? (Evidence means empirical studies of any kind, the body of evidence could be one study; SR may be associated with a guideline)

Answer NO if any:
*Evidence is about something other than what is measured
*Entire body of evidence was not reviewed (just selected studies)
*Expert opinion
*Won't be studied (e.g., "document" diagnosis)
*Distal process step is not the specific focus of the evidence (e.g., monitor BP each visit, when evidence is about treatment of hypertension or relationship to mortality)

4. Is a summary of the quantity, quality, and consistency (QQC) of the body of evidence from a SR provided in the submission form?

A SR is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

Answer NO if:
*QQC not submitted (even if available)
*Details of SR not available (i.e., QQC, evidence tables)

5a. **Does the SR conclude:**
*Quantity:Mod/High; Quality:High; Consistency:High (Table 1)
*High certainty that the net benefit is substantial (e.g., USPSTF-A)
*High quality evidence that benefits clearly outweigh undesirable effects (e.g., GRADE-Strong)
*If measuring inappropriate care, Mod/Hi certainty of no net benefit or harm outweighs benefit (USPSTF-D)

5b. **Does the SR conclude:**
*Quantity:Low-High; Quality:Mod; Consistency:Mod/High (Table 1)
*Moderate certainty that the net benefit is substantial OR moderate-high certainty the net benefit is moderate (e.g., USPSTF-B)

5c. **Does the SR conclude:**
*Consistency:Low; controversial
*Moderate/high certainty that: the net benefit is small (e.g., USPSTF C); OR no net benefit, or harm outweighs benefit (USPSTF-D)
*Low quality evidence, desirable/ undesirable effects closely balanced, uncertainty in preference or use of resources (e.g., GRADE-Weak)

RATE AS LOW
NO *(without QQC from SR, moderate is highest potential rating)*

6. **Does the grade for the evidence or recommendation indicate:**
*High quality evidence (e.g., Table 1 - Quant:Mod/Hi; Qual:Hi; Consist:Hi; USPSTF - High certainty; GRADE-High quality)
*Strong recommendation (e.g., GRADE -Strong; USPSTF-A)

Answer NO if:
*No grading of evidence and summary of QQC not provided

7. Are there, OR could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcome or process?

8. 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?

9. 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?

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 (Appendix B) included:

- Lack of understanding of differences between testing using patient level data versus testing using the computed performance 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 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 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 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 scores are what are used to make conclusions about the quality of care provided. The question is whether the performance 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 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.

The CSAC is interested in receiving 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.

Proposed Guidance on Evaluating Reliability and Validity – Algorithms 2 and 3

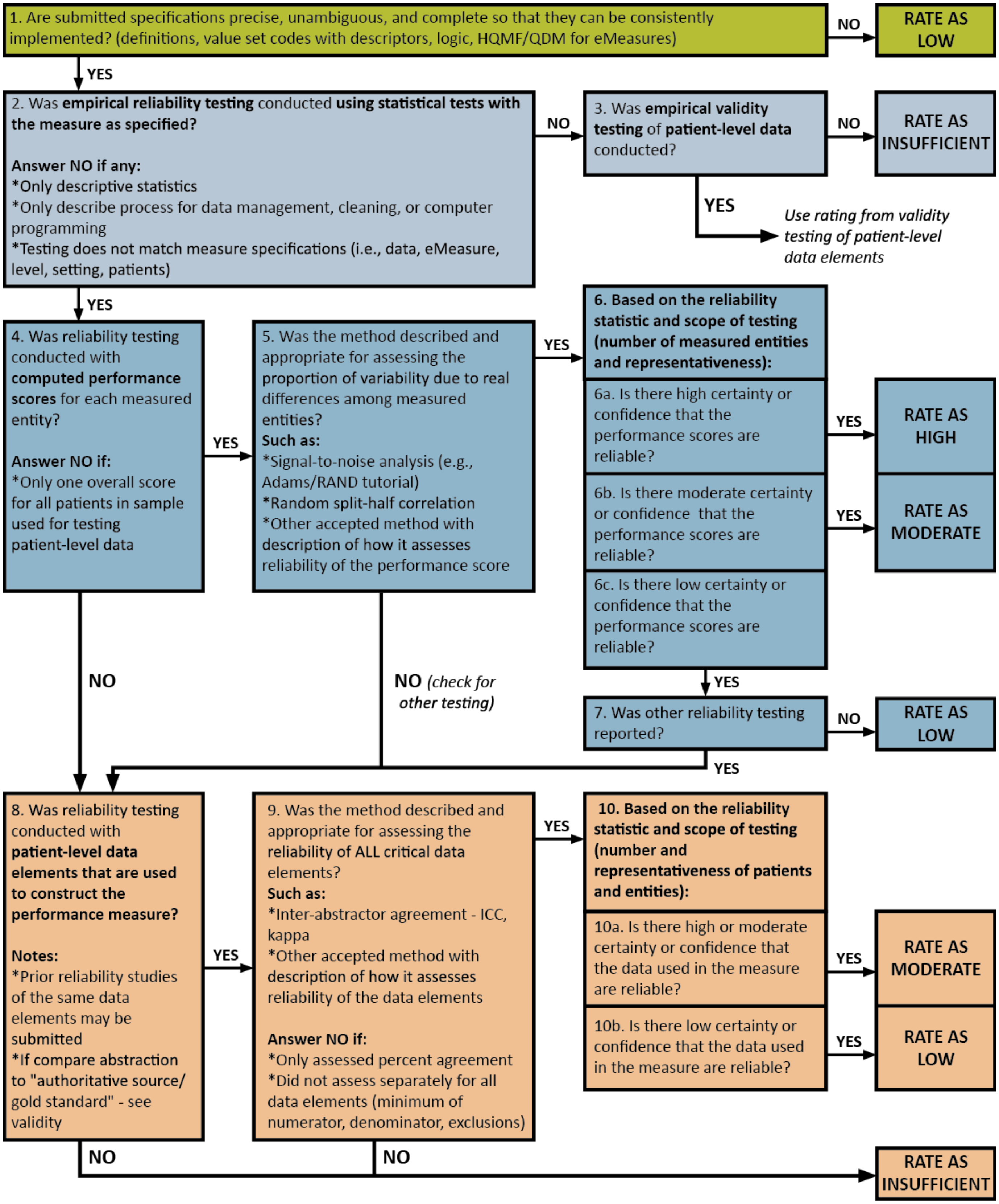
Algorithms 2 and 3 present modified approaches to guide steering committee evaluation of the reliability (Algorithm 2) and validity (algorithm 3) for all measures (including eMeasures). They are consistent with the prior guidance (Appendix B) but are intended to clarify and promote greater consistency 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 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 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.

Algorithm 2. Guidance for Evaluating Reliability



1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? (definitions, value set codes with descriptors, logic, HQMF/QDM for eMeasures)

NO → RATE AS LOW

YES ↓

2. Was **empirical reliability testing** conducted **using statistical tests with the measure as specified?**

Answer NO if any:
*Only descriptive statistics
*Only describe process for data management, cleaning, or computer programming
*Testing does not match measure specifications (i.e., data, eMeasure, level, setting, patients)

NO → 3. Was **empirical validity testing** of **patient-level data** conducted?

NO → RATE AS INSUFFICIENT

YES → *Use rating from validity testing of patient-level data elements*

2. YES ↓

4. Was reliability testing conducted with **computed performance scores** for each measured entity?

Answer NO if:
*Only one overall score for all patients in sample used for testing patient-level data

YES → 5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities?
Such as:
*Signal-to-noise analysis (e.g., Adams/RAND tutorial)
*Random split-half correlation
*Other accepted method with description of how it assesses reliability of the performance score

YES → **6. Based on the reliability statistic and scope of testing (number of measured entities and representativeness):**

6a. Is there high certainty or confidence that the performance scores are reliable? YES → RATE AS HIGH

6b. Is there moderate certainty or confidence that the performance scores are reliable? YES → RATE AS MODERATE

6c. Is there low certainty or confidence that the performance scores are reliable? YES ↓

7. Was other reliability testing reported? NO → RATE AS LOW

7. YES ↓ (to 8)

4. NO ↓ (to 8)

5. **NO** *(check for other testing)* ↓ (to 8)

8. Was reliability testing conducted with **patient-level data elements that are used to construct the performance measure?**

Notes:
*Prior reliability studies of the same data elements may be submitted
*If compare abstraction to "authoritative source/ gold standard" - see validity

YES → 9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
Such as:
*Inter-abstractor agreement - ICC, kappa
*Other accepted method with description of how it assesses reliability of the data elements

Answer NO if:
*Only assessed percent agreement
*Did not assess separately for all data elements (minimum of numerator, denominator, exclusions)

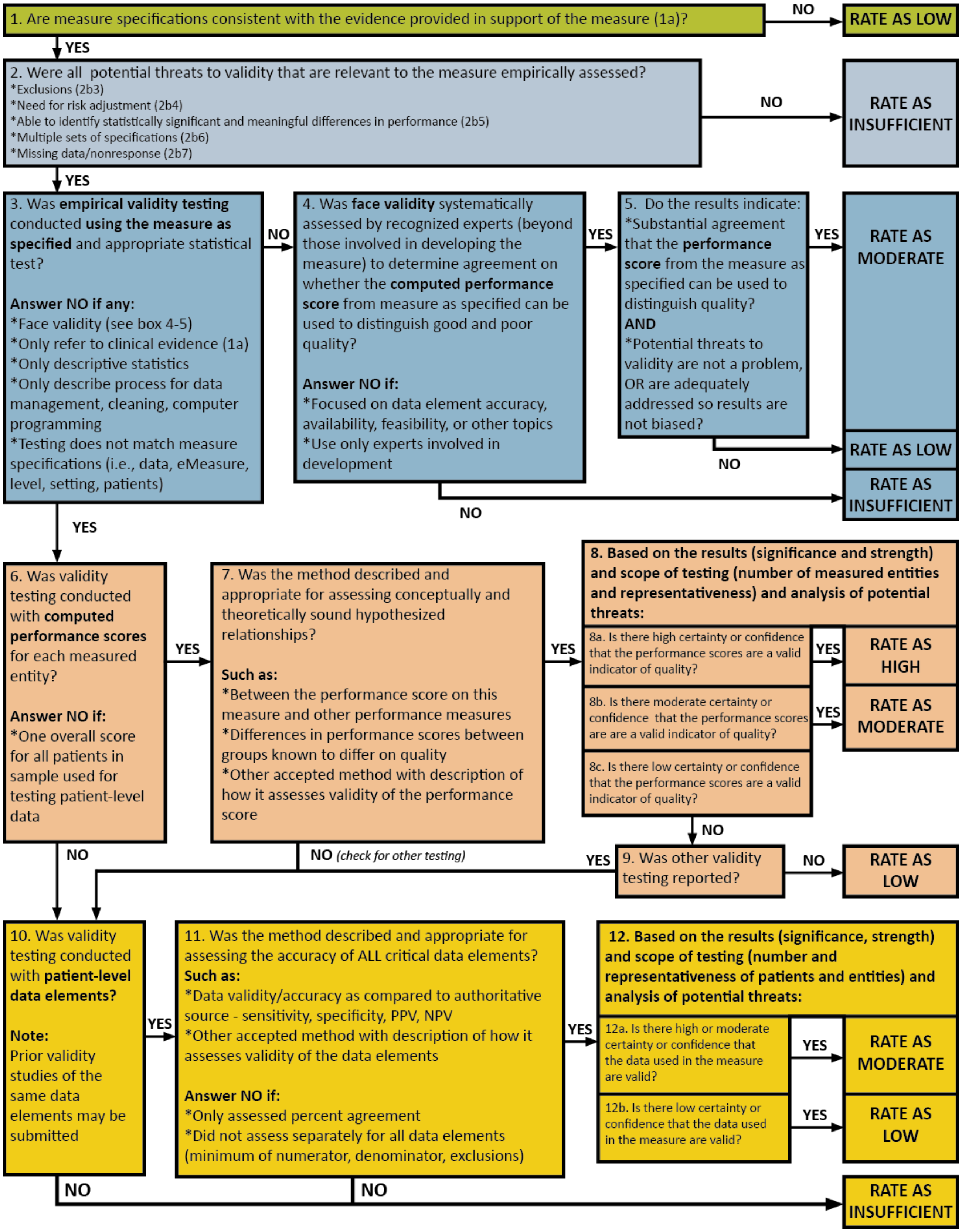
YES → **10. Based on the reliability statistic and scope of testing (number and representativeness of patients and entities):**

10a. Is there high or moderate certainty or confidence that the data used in the measure are reliable? YES → RATE AS MODERATE

10b. Is there low certainty or confidence that the data used in the measure are reliable? YES → RATE AS LOW

8. NO / 9. NO → RATE AS INSUFFICIENT

Algorithm 3. Guidance for Evaluating Validity



1. Are measure specifications consistent with the evidence provided in support of the measure (1a)?

NO → RATE AS LOW

YES ↓

2. Were all potential threats to validity that are relevant to the measure empirically assessed?
*Exclusions (2b3)
*Need for risk adjustment (2b4)
*Able to identify statistically significant and meaningful differences in performance (2b5)
*Multiple sets of specifications (2b6)
*Missing data/nonresponse (2b7)

NO → RATE AS INSUFFICIENT

YES ↓

3. Was **empirical validity testing** conducted **using the measure as specified** and appropriate statistical test?

Answer NO if any:
*Face validity (see box 4-5)
*Only refer to clinical evidence (1a)
*Only descriptive statistics
*Only describe process for data management, cleaning, computer programming
*Testing does not match measure specifications (i.e., data, eMeasure, level, setting, patients)

NO → 4

4. Was **face validity** systematically assessed by recognized experts (beyond those involved in developing the measure) to determine agreement on whether the **computed performance score** from measure as specified can be used to distinguish good and poor quality?

Answer NO if:
*Focused on data element accuracy, availability, feasibility, or other topics
*Use only experts involved in development

YES → 5

NO → RATE AS INSUFFICIENT

5. Do the results indicate:
*Substantial agreement that the **performance score** from the measure as specified can be used to distinguish quality?
AND
*Potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

YES → RATE AS MODERATE

NO → RATE AS LOW

3. YES ↓

6. Was validity testing conducted with **computed performance scores** for each measured entity?

Answer NO if:
*One overall score for all patients in sample used for testing patient-level data

YES → 7

NO → 10

7. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

Such as:
*Between the performance score on this measure and other performance measures
*Differences in performance scores between groups known to differ on quality
*Other accepted method with description of how it assesses validity of the performance score

YES → 8

NO *(check for other testing)* → 10

8. Based on the results (significance and strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats:

8a. Is there high certainty or confidence that the performance scores are a valid indicator of quality?

YES → RATE AS HIGH

8b. Is there moderate certainty or confidence that the performance scores are are a valid indicator of quality?

YES → RATE AS MODERATE

8c. Is there low certainty or confidence that the performance scores are a valid indicator of quality?

NO ↓

9. Was other validity testing reported?

YES → 10

NO → RATE AS LOW

10. Was validity testing conducted with **patient-level data elements?**

Note:
Prior validity studies of the same data elements may be submitted

YES → 11

NO → RATE AS INSUFFICIENT

11. Was the method described and appropriate for assessing the accuracy of ALL critical data elements?
Such as:
*Data validity/accuracy as compared to authoritative source - sensitivity, specificity, PPV, NPV
*Other accepted method with description of how it assesses validity of the data elements

Answer NO if:
*Only assessed percent agreement
*Did not assess separately for all data elements (minimum of numerator, denominator, exclusions)

YES → 12

NO → RATE AS INSUFFICIENT

12. Based on the results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats:

12a. Is there high or moderate certainty or confidence that the data used in the measure are valid?

YES → RATE AS MODERATE

12b. Is there low certainty or confidence that the data used in the measure are valid?

YES → RATE AS LOW

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 following guidance first addresses the criteria for endorsement of eMeasures and then offers a proposed optional alternative pathway for those eMeasures that do not meet the requirements for reliability and validity testing. As described below, the proposed alternative pathway is NOT time-limited endorsement and also is NOT the previously piloted two-stage CDP, but it does contain a few of the elements of those efforts.

Clarification of Requirements for Endorsing eMeasures

The following is a consolidation and clarification of 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). 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 does not support all aspects of a particular measure construct, those may be specified outside HQMF. Please contact NQF staff to discuss format for measure specifications*.
- 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. (See Appendix C for feasibility criteria and example scorecard).
- 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

- All measures (including eMeasures) must be tested for reliability and validity using the data source that is 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 information provided on the data 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 the standardized data element. 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 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 numerator, denominator, and exclusions 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 different EHR systems each with three sites** (total of 9 sites). This requirement is consistent with ONC's [Office of the National Coordinator for Health Information Technology] requirement. Given the proposed optional path of approval for trial use, the HITAC subcommittee agreed that for NQF endorsement, this should be the minimum requirement.
- The following subcriteria under Scientific Acceptability of Measure Properties also apply to eMeasures.
    - Exclusion analysis (2b3). If exclusions 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 scores should be computed for all accountable entities for which you have eMeasure data (not just those on which validity testing was conducted) and analyzed to identify differences in performance.

- Comparability of performance scores if specified for multiple data sources (2b6) (e.g., EHRs, claims). If a performance measure is specified for more than one data source, it should be tested with each. Unless empirical analyses demonstrate comparability of scores computed, assume noncomparability and submit as separate measures. The measures specified for different data sources will be evaluated as competing measures 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.

The HITAC and CSAC are interested in receiving comments on the minimum number of testing sites when conducting validity testing of the data elements for eMeasures and whether a similar requirement should apply to all measures.

Proposed Approval for Trial Use for eMeasures

This optional path of ***approval for trial use*** is intended for eMeasures that are ready for implementation but cannot yet be adequately tested to meet NQF endorsement criteria. For such eMeasures, NQF proposes to utilize the multi-stakeholder consensus process to evaluate and approve eMeasures for trial use that address important areas for performance measurement and quality improvement, though they may not have the requisite testing needed for NQF endorsement. These eMeasures must be assessed to be technically acceptable for implementation. The goal of approving eMeasures for trial use is to promote implementation and the ability to conduct more robust reliability and validity testing that can take advantage of the clinical data in EHRs.

Approval for trial use is NOT time-limited endorsement as it carries no endorsement label. Also, this is not a required two-stage review process: eMeasures that meet endorsement criteria do not need to first go through an approval for trial use.

To be clear, eMeasures that are approved by NQF for trial use would differ from eMeasures that are endorsed.

> *NQF Endorsement* means that the eMeasure has been judged to meet all NQF evaluation criteria and is suitable for use in accountability applications as well as performance improvement.

> *NQF Approval for Trial Use* means the eMeasure has been judged to meet the criteria indicating it is ready to be implemented in real-world settings to generate the data required to assess reliability and validity in the future. It also could be used for internal performance improvement. However, it has not yet been judged to meet all the criteria indicating it is suitable for use in accountability applications.

Criteria for Approval for Trial Use

- Such measures will be considered Approved for Trial Use, NOT Endorsed
- When sufficient data have been accumulated for adequate reliability and validity testing, the eMeasure can be submitted to NQF for potential endorsement (not all may progress to endorsement).
- The following are the proposed requirements for Approval for Trial Use:

- Must be eMeasures, meaning the measures are specified in the accepted standard of HQMF format, and must use the Quality Data Model (QDM). 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 does not support all aspects of a particular measure construct, those may be specified outside HQMF. Please contact NQF staff to discuss format for measure specifications*.)
- Must use value sets vetted through the National Library of Medicine's Value Set Authority Center. This will help ensure appropriate use of codes and code systems and will help minimize value set harmonization issues in submitted eMeasures.
- Must meet all criteria under Importance to measure and report (clinical evidence, performance gap, priority).
- The feasibility assessment must be completed.
- Results from testing with a simulated (or test) data set demonstrate that the QDM and HQMF are used appropriately and that the measure logic performs as expected.
- There is a plan for use and discussion of how the measure will be useful for accountability and improvement.

The CSAC and HITAC are especially interested in receiving comments about approval for trial use.

Appendix A: Current Criteria and Guidance related to Clinical Evidence

1.Evidence, Performance Gap, and Priority (Impact)—Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.*** **Yes□ No□**

1a. Evidence to Support the Measure Focus Yes□ No□

Quantity: H□ M□ L□ I □ Quality: H□ M□ L□ I □ Consistency: H□ M□ L□ I □

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

AND

1b. Performance Gap H□ M□ L□ I □

Demonstration of quality problems and opportunity for improvement, i.e., data [7] demonstrating

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

AND

1c. High Priority (previously referred to as High Impact) **H□ M□ L□ I □**
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;

OR

- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).
- For patient-reported outcomes, there is evidence that the target population values the PRO and finds it meaningful.

1d. For composite performance measures, the following must be explicitly articulated and logical:

1d1. The quality construct, including the overall area of quality; included component measures; and the

relationship of the component measures to the overall composite and to each other; and
1d2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and
1d3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).
7. Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, or data from pilot testing or implementation of the proposed measure. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

Table A-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.

Table A-2: Evaluation of Subcriterion 1a Based on the Quantity, Quality, and Consistency of the Body of Evidence

QUANTITY OF BODY OF EVIDENCE	**QUALITY** OF BODY OF EVIDENCE	**CONSISTENCY** OF RESULTS OF BODY OF EVIDENCE	**PASS SUBCRITERION 1A**
Moderate-High	Moderate-High	Moderate-High	Yes
Low	Moderate-High	Moderate (if only one study, high consistency not possible)	Yes, but only if it is judged that additional research is unlikely to change conclusion that benefits to patients outweigh harms; otherwise, No
Moderate-High	Low	Moderate-High	Yes, but only if it is judged that potential benefits to patients clearly outweigh potential harms; otherwise, No
Low-Moderate-High	Low-Moderate-High	Low	No
Low	Low	Low	No
Exception to Empirical Body of Evidence for Health Outcome For a health outcome measure: A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service			Yes, if it is judged that the rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service
Potential Exception to Empirical Body of Evidence for Other Types of Measures If there is no empirical evidence, expert opinion is systematically assessed with agreement that the benefits to patients greatly outweigh potential harms.			Yes, but only if it is judged that potential benefits to patients clearly outweigh potential harms; otherwise, No

Appendix B: Current Criteria and Guidance related to Reliability and Validity Testing

2. Reliability and Validity—Scientific Acceptability of Measure Properties
Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria***.
Yes□ No□

2a. Reliability H□ M□ L□ I □
2a1. The measure is well defined and precisely specified [8] so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the quality data model (QDM). [9]

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b. Validity H□ M□ L□ I □
2b1. The measure specifications [8] are consistent with the evidence presented to support the focus of measurement under criterion 1a. The measure is specified to capture the most inclusive target population indicated by the evidence, and exclusions are supported by the evidence.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. . For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful [16] differences in performance;
OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2c. Disparities *(Disparities should be addressed under subcriterion 1b)*
If disparities in care have been identified, measure specifications, scoring, and analysis allow for identification of disparities through stratification of results (e.g., by race, ethnicity, socioeconomic status, gender);

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate the following:
2d1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2d2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(*if not conducted or results not adequate, justification must be submitted and accepted)*

Notes
8. Measure specifications include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation.
Specifications for **PRO-PMs** also include: specific PROM(s); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and calculation of response rates to be reported with the performance measure results.
Specifications for **composite performance measures** include: component measure specifications (unless individually endorsed); aggregation and weighting rules; handling of missing data; standardizing scales across component measures; required sample sizes.
9. Some eMeasures may have aspects that cannot be specified in HQMF or the QDM (e.g., risk adjustment, composite aggregation and weighting rules) and should be specified in HQMF and QDM to the extent possible, with the additional specifications and an explanation why cannot be represented in HQMF or QDM. eMeasure specifications include data type from the QDM, value sets and attributes, measure logic, original source of the data and recorder.
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. 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). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable

to stratify measures by race and socioeconomic status rather than to adjust out the differences.
16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Table B-1: Evaluation Ratings for Reliability and Validity

RATING	RELIABILITY	VALIDITY
High	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.; **AND** Empirical evidence of reliability of **BOTH** data elements **AND** computed performance measure score within acceptable norms: • Data element: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity was demonstrated*; **AND** • Performance measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1a) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of **BOTH** data elements **AND** computed performance measure score within acceptable norms: • Data element: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **AND** • Performance measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased
Moderate	All measure specifications are unambiguous as noted above **AND** Empirical evidence of reliability within acceptable norms for either critical data elements **OR** performance measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either critical data elements **OR** performance measure score as noted above; **OR** Systematic assessment of face validity of performance measure score as a quality indicator explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased

RATING	RELIABILITY	VALIDITY
Low	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are ambiguous with potential for confusion in identifying who is included and excluded from the target population, or the event, condition, or outcome being measured; or how to compute the score, etc.; **OR** Empirical evidence (using appropriate method and scope) of unreliability for either data elements **OR** performance measure score, i.e., statistical results outside of acceptable norms	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** performance measure score, i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias results
Insufficient Evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above); **OR** Threats to validity as noted above are likely and are NOT empirically assessed

Table B-2: Evaluation of Reliability and Validity of eMeasures

RATING	RELIABILITY DESCRIPTION AND EVIDENCE	VALIDITY DESCRIPTION AND EVIDENCE
High	Specified in HQMF and QDM and all specifications are unambiguous[+]; **AND** Empirical evidence of reliability of both data element **AND** computed performance measure score within acceptable norms: • Data element: reliability (repeatability) assured with computer programming—**must test data element validity** **AND** • Performance measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1a) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of both data elements **AND** computed performance measure score within acceptable norms: • Data element: validity demonstrated by analysis of agreement between data elements obtained using the eMeasure aspecifications and data elements manually abstracted from the entire EHR with statistical results within acceptable norms; **OR** complete agreement between data elements and computed measure scores obtained by applying the eMeasure specifications to a simulated test EHR data set with known values for the critical data elements; **AND** • Performance measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased
Moderate	Specified in HQMF and QDM and all specifications are unambiguous[+] and include only data elements from the QDM;* **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability within acceptable norms for either data elements **OR** performance measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either data elements **OR** performance measure score as noted above; **OR** Systematic assessment of face validity of performance measure score as a quality indicator explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased

RATING	RELIABILITY DESCRIPTION AND EVIDENCE	VALIDITY DESCRIPTION AND EVIDENCE
Low	One or more eMeasure specifications are ambiguous[+] or do not use HQMF and QDM*; **OR** Empirical evidence of unreliability for either data elements **OR** performance measure score—i.e., statistical results outside of acceptable norms	The EHR measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** performance measure score— i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias results
Insufficient evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above) **OR** Threats to validity as noted above are likely and are NOT empirically assessed

[+] Specifications are considered unambiguous if they are likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.

*HQMF and QDM (formerly called the QDS) should be used when available. When quality data elements are needed but are not yet available in the QDM, they will be considered for addition to the QDM.

Table B-3: Evaluation of Scientific Acceptability of Measure Properties Based on Reliability and Validity Ratings

VALIDITY RATING	RELIABILITY RATING	PASS *SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES* FOR INITIAL ENDORSEMENT*	
High	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

*A measure that does not pass the criterion of *Scientific Acceptability of Measure Properties* would not be recommended for endorsement.

Table B-4: Scope of Testing Required at the Time of Review for Endorsement Maintenance

	FIRST ENDORSEMENT MAINTENANCE REVIEW	SUBSEQUENT REVIEWS
Reliability	**Measure In Use** • Analysis of data from entities whose performance is measured • Reliability of measure scores (e.g., signal to noise analysis) **Measure Not in Use** • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that reliability achieved a high rating
Validity	**Measure in Use** • Analysis of data from entities whose performance is measured • Validity of measure score for making accurate conclusions about quality • Analysis of threats to validity **Measure Not in Use** • Expanded testing in terms of scope (number of entities/patients) and/or levels (data elements/measure score)	Could submit prior testing data, if results demonstrated that validity achieved a high rating

Appendix C: Current Criteria and Guidance related to Feasibility

3. Feasibility:
Extent to which the specifications, including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
H□ **M**□ **L**□ **I** □

3a. For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order). **H**□ **M**□ **L**□ **I** □

3b. The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified. **H**□ **M**□ **L**□ **I** □

3c. Demonstration that the data collection strategy (e.g., data source/availability, timing, frequency, sampling, patient-reported data, patient confidentiality, [17] costs associated with fees/licensing for proprietary measures or elements such as risk model, grouper, instrument) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic [18] and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.
H□ **M**□ **L**□ **I** □

Note
17. All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.
18. The feasibility assessment uses a standard score card or a fully transparent alternative that includes at a minimum: a description of the assessment, feasibility scores for all data elements, and explanatory notes for all data element components scoring a "1" (lowest rating); measure logic can be executed; with rationale and plan for addressing feasibility concerns.

Guidance on Evaluating Feasibility

Table C-1: Generic Scale for Rating Feasibility Subcriteria

RATING	DEFINITION
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

Table C-2. Data Element Feasibility Scorecard

DATA ELEMENT:			
Measure Title:			
Data element definition:			
Who performed the assessment:			
Type of setting or practice, i.e., solo practice, large group, academic hospital, safety net hospital, integrated system:			
EHR system used:			
	Current (1-3)	Future* (1-3)	Comments
Data Availability – Is the data readily available in structured format? Scale: 3 – Data element exists in structured format in this EHR. [2] – Not defined as this time. Hold for possible future use. 1 – Data element is not available in structured format in this EHR.			
***Data Accuracy* –** Is the information contained in the data element correct? Are the data source and recorder specified? Scale: 3 – The information is from the most authoritative source and/or is highly likely to be correct. (e.g., laboratory test results transmitted directed from the laboratory information system into the EHR). 2 – The information may not be from the most authoritative source and/or has a moderate likelihood of being correct. (e.g., self-report of a vaccination). 1 – The information may not be correct. (e.g., a check box that indicates medication reconciliation was performed).			
***Data Standards* –** Is the data element coded using a nationally accepted terminology standard? Scale: 3 – The data element is coded in nationally accepted terminology standard. 2 – Terminology standards for this data element are currently available, but is not consistently coded to standard terminology in the EHR, or the EHR does not easily allow such coding. 1 – The EHR does not support coding to the existing standard.			

DATA ELEMENT:			
***Workflow* –** To what degree is the data element captured during the course of care? How does it impact the typical workflow for that user? Scale: 3 – The data element is routinely collected as part of routine care and requires no additional data entry from clinician solely for the quality measure and no EHR user interface changes. Examples would be lab values, vital signs, referral orders, or problem list entry. 2 – Data element is not routinely collected as a part of routine care and additional time and effort over and above routine care is required, but perceived to have some benefit. 1 – Additional time and effort over and above routine care is required to collect this data element without immediate benefit to care			
DATA ELEMENT FEASIBILITY SCORE			

*For data elements that score low on current feasibility, indicate the anticipated feasibility score in 3-5 years based on a projection of the maturation of the EHR, or maturation of its use.

Appendix D: 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
BoozAllenHamilton, 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