NATIONAL QUALITY FORUM

Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement

Effective April 2015

Contents

Changes from the 2013 Criteria and Guidance

This document updates the 2013 Measure Evaluation Criteria and Guidance. Additionally, the document consolidates several NQF documents pertaining to the criteria and evaluation into a single document.

- Subcriterion 1a. Includes additional guidance for patient-reported outcome measures and appropriate use measures.
- Subcriterion 1b. Opportunity for improvement. Guidance has been expanded to discuss "topped out" measures and NQF's policy for Inactive Endorsement with Reserve Status.
- Subcriterion 1c. High Priority was removed as an evaluation criterion by the Consensus Standards Approval Committee (CSAC).
- Subcriterion 2b4 Risk adjustment: In late 2014, the NQF Board of Directors approved, for a trial period, a change in the policy that prohibited the use of sociodemographic factors in statistical risk models. During the trial period, risk-adjusted measures submitted to NQF for evaluation should be submitted with analysis of both clinical and non-clinical SDS factors considered for inclusion in risk adjustment models. Details on the SDS Trial Period are included.
- Guidance for evaluating eMeasures has been included. This guidance is updated from Review and Update of Guidance for Evaluating Evidence and Measure Testing - Technical Report (October 2013).

Introduction

This document contains the measure evaluation criteria as well as additional guidance for evaluating measures based on the criteria. Additional information is available in detailed reports that can be accessed through NQF's Measure Evaluation webpage.

Conditions for Consideration

Several conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards. **If any of the conditions are not met, the measure will not be accepted for consideration.**

A. The measure is in the public domain or a measure steward agreement is signed.

B. The measure owner/steward verifies that there is an identified responsible entity and a process to maintain and update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every 3 years.

C. The intended use of the measure includes both accountability applications[1] (including public reporting) and performance improvement to achieve high-quality, efficient healthcare.

D. The measure is fully specified and tested for reliability and validity. [2]

E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate.

F. The requested measure submission information is complete and responsive to the questions so that all the information needed to evaluate all criteria is provided.

Note

1. Accountability applications are the use of performance results about identifiable, accountable entities to make judgments and decisions as a consequence of performance, such as reward, recognition, punishment, payment, or selection (e.g., public reporting, accreditation, licensure, professional certification, health information technology incentives, performance-based payment, network inclusion/exclusion). **Selection** is the use of performance results to make or affirm choices regarding providers of healthcare or health plans.

2. An eMeasure that has not been tested sufficiently to meet endorsement criteria may be eligible for Approval for Trial Use. Time-limited endorsement is no longer available.

Criteria for Evaluation

If all conditions for consideration are met, measures are evaluated for their suitability based on standardized criteria in the following order: *Importance to Measure and Report*, *Scientific Acceptability of Measure Properties*, *Feasibility, Usability and Use*, and *Related and Competing Measures*. Not all acceptable measures will be equally strong on each set of criteria. The assessment of each criterion is a matter of degree. However, if a measure is not judged to have met minimum requirements for *Importance to Measure and Report* or *Scientific Acceptability of Measure Properties*, it cannot be recommended for endorsement and will not be evaluated against the remaining criteria. These criteria apply to all performance measures (including outcome and resource use measures, PRO-PMs, composite performance measures, eMeasures), except where indicated for a specific type of measure.

For **composite performance measures**, the following subcriteria apply to each of the component measures: 1a; 1b (also composite); 2b3 (also composite); 2b4; 2b6; 4c (also composite); 5a and 5b (also composite).

1.Evidence and Performance Gap, Importance to Measure and Report

Extent to which the specific measure focus is evidence-based and important to making significant gains in healthcare quality 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

Use Algorithm 1 and Table 2 to rate this criterion. H☐ M☐ L☐ I ☐

The measure focus is evidence-based, demonstrated as follows (Table 1):

- 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 is required for the quality component but not required for the resource use component. **(**Measures of efficiency combine the concepts of resource use and quality.
- Patient-reported outcome-based performance measures (PRO-PMs): in addition to evidence required for any outcome measure, evidence should demonstrate that the target population values the measured PRO and finds it meaningful (see Table 13).
- Measures incorporating Appropriate Use Criteria: NQF's guidance for evidence for measures in general, and specifically those based on clinical practice guidelines, apply to measures based on appropriateness criteria as well. (see Guidance on Evaluating Evidence for Appropriate Use Measures).

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 1 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).

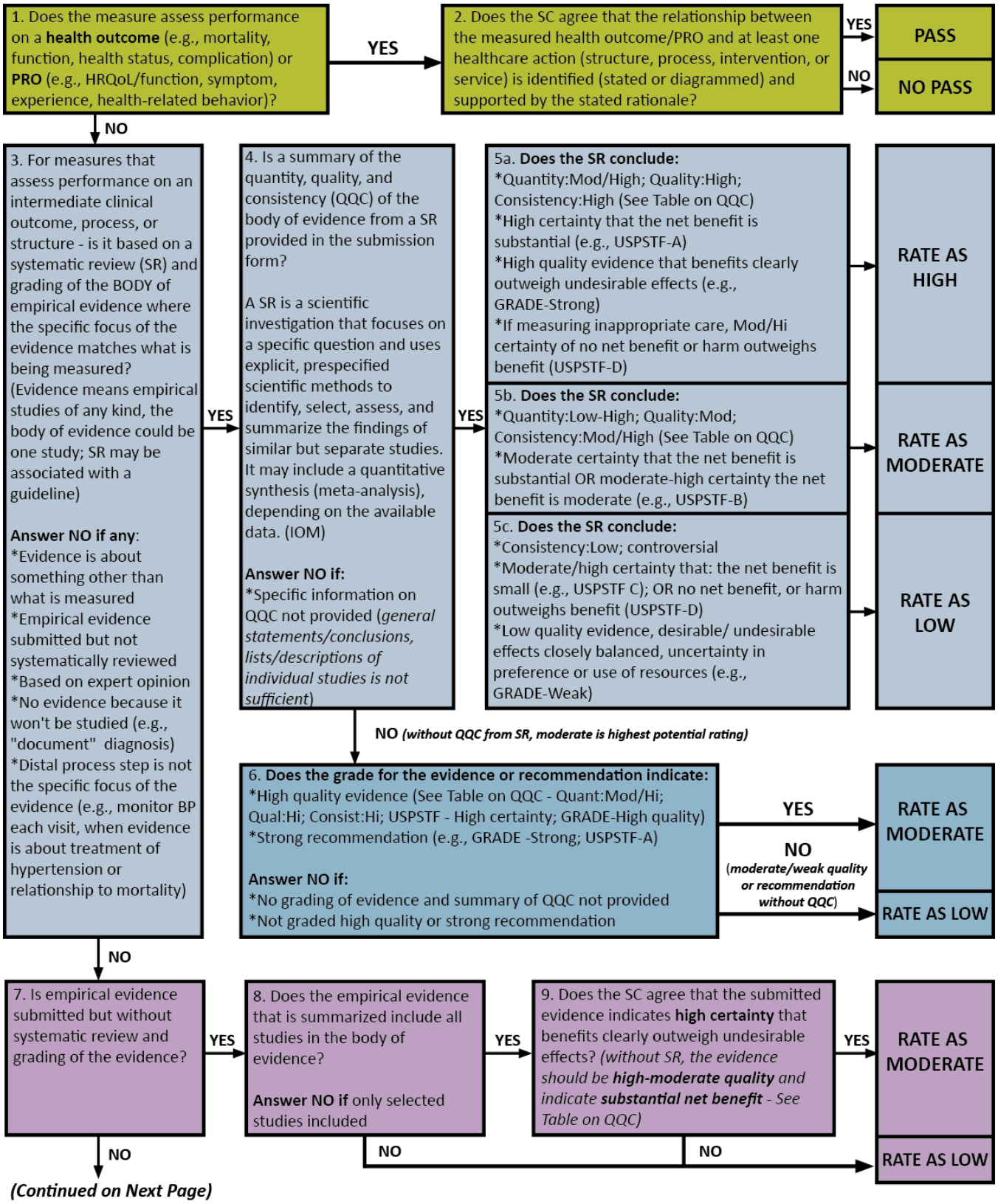
Guidance on Evaluating Importance to Measure and Report

Table 1. Evidence to Support the Focus of Measurement

Type of Measure	Evidence	Example of Measure Type and Evidence to Be Addressed
Health Outcome An outcome of care is the health status of a patient (or change in health status) resulting from healthcare—desirable or adverse. In some situations, resource use may be considered a proxy for a health state (e.g., hospitalization may represent deterioration in health status). **Patient-reported outcomes** include health-related quality of life/functional status, symptom/ symptom burden, experience with care, health-related behavior	A rationale supports the relationship of the health outcome to at least 1 healthcare structure, process, intervention, or service.	**#0230** Acute myocardial Infarction 30-day mortality Survival is a goal of seeking and providing treatment for AMI. **Rationale** healthcare processes/ interventions (aspirin, reperfusion) lead to decreased mortality/ increased survival **#0171** Acute care hospitalization (risk-adjusted) [of home care patients] Improvement or stabilization of condition to remain at home is a goal of seeking and providing home care services. **Rationale** healthcare processes (medication reconciliation, care coordination) lead to decreased hospitalization of patients receiving home care services **#0140** Ventilator-associated pneumonia for ICU and high-risk nursery (HRN) patients Avoiding harm from treatment is a goal when seeking and providing healthcare. **Rationale** healthcare processes (ventilator bundle) lead to decreased ventilator acquired pneumonia **#0711** Depression remission at 6 months Relief of symptoms is a goal of seeking and providing healthcare services. **Rationale:** healthcare processes (use of antidepressants, psychotherapy) lead to decreased symptoms of depression #0166 HCAHPS experience with communication with doctors (assuming demonstration this is of value to patients) **Rationale:** healthcare practices (response time, respect, attention, explanation) leads to better experience with physician communication

Type of Measure	Evidence	Example of Measure Type and Evidence to Be Addressed
Intermediate Clinical Outcome An intermediate outcome is a change in physiologic state that leads to a longer-term health outcome.	Quantity, quality, and consistency of a body of evidence that the measured intermediate clinical outcome leads to a desired health outcome.	**#0059** Hemoglobin A1c management [A1c > 9] **Evidence** that hemoglobin A1c level leads to health outcomes (e.g., prevention of renal disease, heart disease, amputation, mortality)
Process A process of care is a healthcare-related activity performed for, on behalf of, or by a patient.	Quantity, quality, and consistency of a body of evidence that the measured healthcare process leads to desired health outcomes in the target population with benefits that outweigh harms to patients. Specific drugs and devices should have FDA approval for the target condition. If the measure focus is on inappropriate use, then quantity, quality, and consistency of a body of evidence that the measured healthcare process does *not* lead to desired health outcomes in the target population.	**#0551** ACE inhibitor/Angiotensin receptor blocker (ARB) use and persistence among members with coronary artery disease at high risk for coronary events **Evidence** that use of ACE-I and ARB results in lower mortality and/or cardiac events **#0058** Inappropriate antibiotic treatment for adults with acute bronchitis **Evidence** that antibiotics are not effective for acute bronchitis
Structure Structure of care is a feature of a healthcare organization or clinician related to its capacity to provide high-quality healthcare.	Quantity, quality, and consistency of a body of evidence that the measured healthcare structure leads to desired health outcomes with benefits that outweigh harms (including evidence for the link to effective care processes and the link from the care processes to desired health outcomes).	**#0190** Nurse staffing hours **Evidence** that higher nursing hours result in lower mortality or morbidity, or lead to provision of effective care processes (e.g., lower medication errors) that lead to better outcomes

Algorithm 1. Guidance for Evaluating the Clinical Evidence



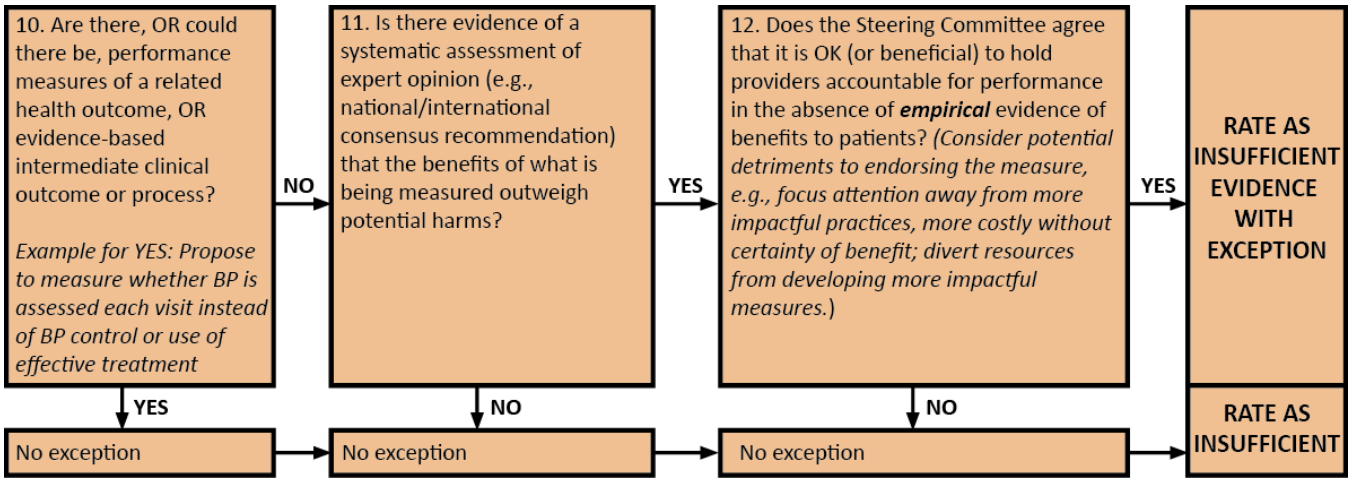

10. Are there, OR could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcome or process?
Example for YES: Propose to measure whether BP is assessed each visit instead of BP control or use of effective treatment
NO
11. Is there evidence of a systematic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

Table 2. 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 1 study, then the estimate of benefits greatly outweighs the estimate of potential harms to patients (1 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 1 study, then estimated benefits do not greatly outweigh harms to patients
Insufficient to Evaluate	• 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.
[b]The suggested number of studies for rating levels of quantity is considered a general guideline.

1.Evidence and Performance Gap -- Importance to Measure and Report (continued)
Extent to which the specific measure focus is evidence-based and important to making significant gains in healthcare quality, 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 (see above)

AND

1b. Performance Gap

Use Table 3 to rate criterion. 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.

When assessing measure performance data for Performance Gap (1b), the following factors should be considered:

- distribution of performance scores;
- number and representativeness of the entities included in the measure performance data;
- data on disparities; and
- size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.

For maintenance of endorsement: If a measure is found to be "topped out" (i.e., does not meet criteria for opportunity for improvement (1b)), the measure will be considered for inactive endorsement with reserve status only. The measure must meet all other criteria, otherwise the measure should not be endorsed. See Inactive Endorsement with Reserve Status policy.

1c. For composite performance measures, the following must be explicitly articulated and logical: **H☐ M☐ L☐ I ☐**

1c1. 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
1c2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and
1c3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

Notes
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 3. Generic Scale for Rating Subcriteria 1b, 1c

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).

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 Use algorithm 2 to rate criterion. 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.

For any measures that use ICD-9 CM codes, ICD-10 CM codes must also be provided. If HHS implements ICD-10 as planned in October 2015, then NQF will no longer accept ICD-9 CM codes for measures after December 31, 2015.

eMeasures should be specified in the Health Quality Measures Format (HQMF) and must use the Quality Data Model (QDM) and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC). [9]

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.

2a2. Reliability testing[10] demonstrates that 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 Use algorithm 3 to rate the criterion. 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**, validity 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 is specified; is based on patient factors (including clinical and sociodemographic risk factors) that influence the measured outcome and are present at start of care;[14,15] and has demonstrated adequate discrimination and calibration.

OR

- rationale/data support no risk adjustment. (See section on SDS Trial Period)

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: H☐ M☐ L☐ I ☐
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)*

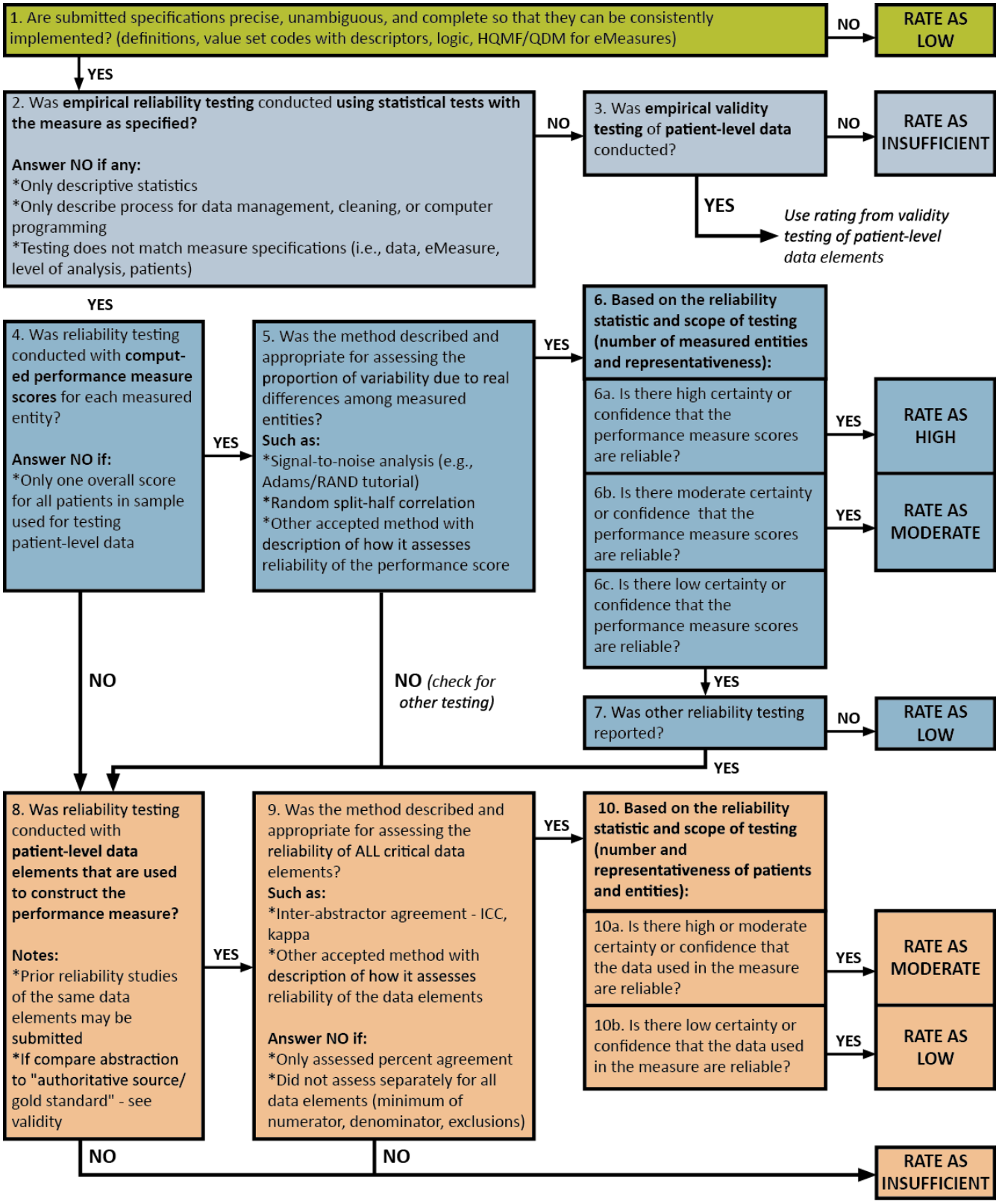
Notes
8. Measure specifications include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation.
9. 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.~~ In late 2014, the NQF Board of Directors approved, for a trial period, a change in the policy that prohibited the use of sociodemographic factors in statistical risk models. During the trial period, risk-adjusted measures submitted to NQF for evaluation may include both clinical and sociodemographic factors in the risk adjustment models. See section on SDS Trial Period.

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 1 percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74% versus 75%) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 versus $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Algorithm 2. Guidance for Evaluating Reliability (including eMeasures)



Algorithm 3. Guidance for Evaluating Validity (including eMeasures)

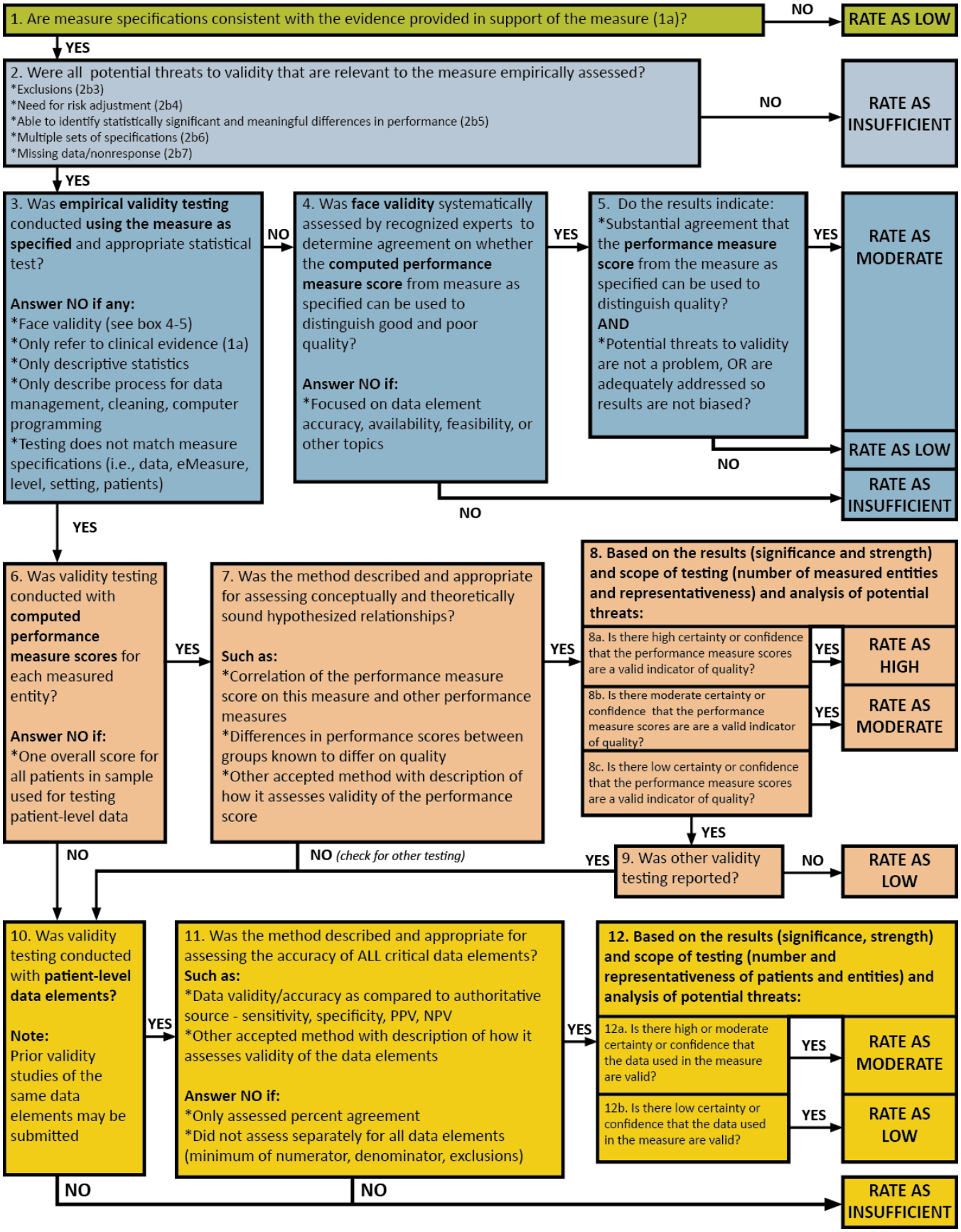


Table 4. Scope of Testing Required at the Time of Review for Endorsement Maintenance

	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 good reliability.
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

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.
Use Table 5 to rate criterion. **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 is required; this feasibility assessment must address the data elements and measure logic and demonstrate that the eMeasure can be implemented or that feasibility concerns can be adequately addressed. The feasibility assessment uses a standard score card or a fully transparent alternative that includes at a minimum: 1)a description of the assessment, feasibility scores for all data elements, and explanatory notes for all data element components scoring a "1" (lowest rating); 2) demonstration that the measure logic can be executed; w and 3) plan for addressing feasibility concerns (see Table 6).
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.

Guidance on Evaluating Feasibility

Table 5. 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 6. Data Element Feasibility Scorecard

Data Element:			
eMeasure 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:			
Component	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.			
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.

4. Usability and Use
Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations. **Use Table 7 to rate criterion H☐ M☐ L☐ I ☐**

4a. Accountability and Transparency
Performance results are used in at least 1 accountability application[1] within 3 years after initial endorsement and are publicly reported[18] within 6 years after initial endorsement (or the data on performance results are available).[19] If not in use at the time of initial endorsement, then a credible plan[20] for implementation within the specified timeframes is provided.
H☐ M☐ L☐ I ☐

AND

4b. Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.[21] If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.
H☐ M☐ L☐ I ☐

AND

4c. The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists). **H☐ M☐ L☐ I ☐**

Notes
18. Transparency is the extent to which performance results about identifiable, accountable entities are *disclosed and available* outside of the organizations or practices whose performance is measured. Maximal transparency is achieved with **public reporting** defined as making comparative performance results about identifiable, accountable entities freely available (or at nominal cost) to the public at large (generally on a public website). *At a minimum, the data on performance results about identifiable, accountable entities are available to the public (e.g., unformatted database)*. The capability to verify the performance results adds substantially to transparency.
19. This guidance is not intended to be construed as favoring measures developed by organizations that are able to implement their own measures (such as government agencies or accrediting organizations) over equally strong measures developed by organizations that may not be able to do so (such as researchers, consultants, or academics). Accordingly, measure developers may request a longer timeframe with appropriate explanation and justification.
20. Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.
21. An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement. Demonstrated progress toward achieving the goal of high-quality, efficient healthcare includes evidence of improved performance and/or increased numbers of individuals receiving high-quality healthcare. Exceptions may be considered with appropriate explanation and justification.

Guidance on Evaluating Usability and Use

Table 7. Generic Scale for Rating Usability and Use 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 8. Key Questions for Evaluating Usability and Use

Subcriteria	Key Questions	Suitable for Endorsement?
4a., 4b, 4c	• Are all 3 subcriteria met? (3a—accountability/transparency, 3b—improvement, and 3c—benefits outweigh any unintended consequences)	If Yes, then the Usability and Use criterion is met, and if the other criteria (Importance to Measure and Report, Scientific Acceptability of Measure Properties, Feasibility) are met, then the measure is **suitable for endorsement**
4a. Accountability/Transparency	• Is it an initial submission with a credible plan for implementation in an accountability application? • Is the measure used in at least 1 accountability application by 3 years? • Are the performance results publicly reported by 6 years (or the data on performance results are available)? If any of the above answers are "No": • What are the reasons (e.g., developer/steward, external factors)? • Is there a credible plan for implementation and public reporting?	If 4a and/or 4b are not met, then the Usability and Use criterion is not met, but the measure **may or not be suitable for endorsement** depending on an assessment of the following: • timeframe (initial submission, 3 years, 6 years, or longer); • reasons for lack of use in accountability application/public reporting (4a) and/or lack of improvement (4b); • credibility of plan for implementation for accountability/public reporting (4a) and/or credibility of rationale for improvement (4b); • strength of the measure in terms of the other three criteria (Importance to Measure and Report, Scientific Acceptability of Measure Properties, and Feasibility); and • strength of competing and related measures to drive improvement. Exceptions to the timeframes for accountability and public reporting (4a) OR demonstration of improvement (4b) require judgment and supporting rationale.
4b. Improvement	• Is it an initial submission with a credible rationale for improvement? • Has improvement been demonstrated (performance trends, numbers of people receiving high-quality, efficient healthcare)? If any of the above answers are "No": • What are the reasons? • Is there a credible rationale describing how the performance results could be used to further the goal of facilitating high-quality, efficient healthcare for individuals or populations? • Is the measure used in quality improvement programs?	
4c. Unintended negative consequences	• Is there **evidence** that unintended negative consequences to individuals or populations outweigh the benefits? For most measures, this will not be applicable and will not be a factor in whether a measure is recommended.	If Yes, then the Usability and Use criterion is not met and the measure is **not suitable for endorsement** regardless of evaluation of 4a and 4b.

5. Comparison to Related or Competing Measures
If a measure meets the above criteria and there are endorsed or new related measures (with either the same measure focus or the same target population) or competing measures (both with the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5a. The measure specifications are harmonized[23] with related measures;
OR
the differences in specifications are justified.

5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
multiple measures are justified.

Note
23. Measure harmonization refers to the standardization of specifications for related measures with the same measure focus (e.g., *influenza immunization* of patients in hospitals or nursing homes); related measures with the same target population (e.g., eye exam and HbA1c for *patients with diabetes*); or definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless differences are justified (e.g., dictated by the evidence). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions. The extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources.

Guidance on Evaluating Related and Competing Measures

Table 9. Related versus Competing Measures

	Same Concepts for Measure Focus—Target Process, Condition, Event, Outcome	Different Concepts for Measure Focus—Target Process, Condition, Event, Outcome
Same target patient population	**Competing measures—Select best measure** from competing measures or justify endorsement of additional measure(s).	**Related measures—Harmonize** on target patient population or justify differences.
Different target patient population	**Related measures—Combine** into 1 measure with expanded target patient population or justify why different harmonized measures are needed.	Neither harmonization nor competing measure issue

Figure 1. Addressing Competing Measures and Harmonization of Related Measures in the NQF Evaluation Process

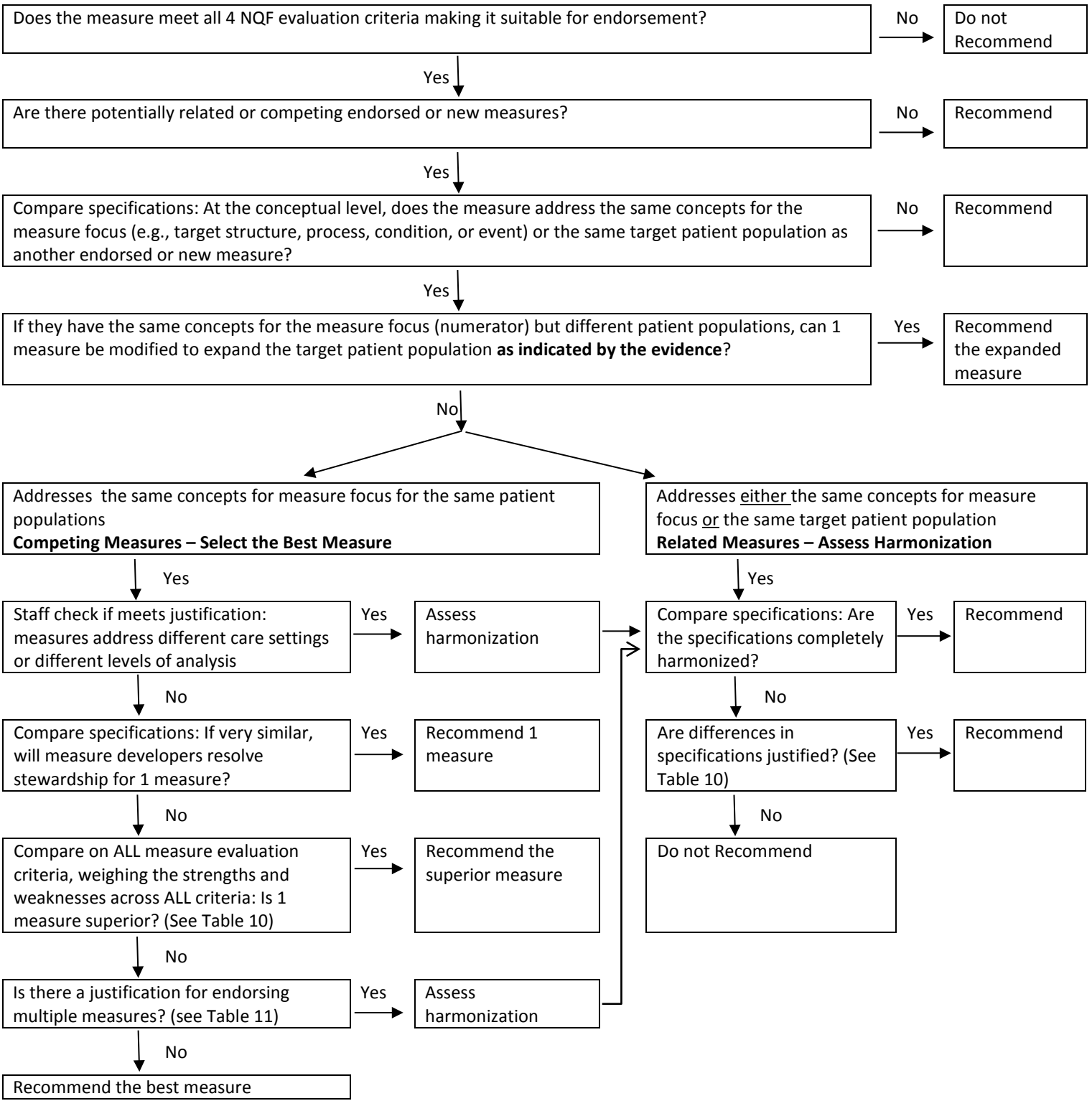


Table 10. Evaluating Competing Measures for Superiority or Justification for Multiple Measures

Steps	Evaluate Competing Measures
1. Determine if need to compare measures for superiority	Work through the steps in the algorithm (Figure 1) to determine if need to evaluate competing measures for superiority (i.e., 2 or more measures address the same concepts for measure focus for the same patient populations)
2. Assess Competing Measures for superiority by weighing the strengths and weaknesses across ALL NQF evaluation criteria	**Because the competing measures have already been determined to have met NQF's criteria for endorsement, the assessment of competing measures must include weighing the strengths and weaknesses across ALL the criteria and involves more than just comparing ratings.** (For example, a decision is not based on just the differences in scientific acceptability of measure properties without weighing the evaluation of importance to measure and report, usability, and feasibility as well.) **Evidence, Performance Gap—Importance to Measure and Report:** Competing measures generally will be the same in terms of the evidence for the focus of measurement (1a) and addressing a high-priority area of healthcare (1b) . However, due to differences in measure construction, they could differ on performance gap. • Compare measures on opportunity for improvement (1b) • **Reliability and Validity—Scientific Acceptability of Measure Properties:** • Compare evidence of reliability (2a1-2a2) • Compare evidence of validity, including threats to validity (2b1-2b7) Untested measures cannot be considered superior to tested measures because there would be no empirical evidence on which to compare reliability and validity. However, a new measure, when tested, could ultimately demonstrate superiority over an endorsed measure and the NQF endorsement maintenance cycles allow for regular submission of new measures. Compare and identify differences in specifications **All else being equal on the criteria and subcriteria**, the preference is for: • Measures specified for the broadest application (target patient population as indicated by the evidence, settings, level of analysis) • Measures that address disparities in care when appropriate **Feasibility:** • Compare the ease of data collection/availability of required data **All else being equal on the criteria and subcriteria**, the preference is for: • Measures based on data from electronic sources • Clinical data from EHRs • Measures that are freely available **Usability and Use:** • Compare evidence of the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement. **All else being equal on the criteria and subcriteria**, the preference is for: • Measures used in at least 1 accountability application • Measures with the widest use (e.g., settings, numbers of entities reporting performance results) • Measures for which there is evidence of progress towards achieving high-quality, efficient healthcare for individuals or populations • The benefits of the measure outweigh any unintended negative consequences to individuals or

Steps	Evaluate Competing Measures
	populations **After weighing the strengths and weaknesses across ALL criteria, identify if 1 measure is clearly superior and provide the rationale based on the NQF criteria.**
3. If a competing measure does not have clear superiority, assess justification for multiple measures	**If a competing measure does not have clear superiority, is there a justification for endorsing multiple measures? Does the added value offset any burden or negative impact?** **Identify the value of endorsing competing measures** Is an additional measure necessary? • to change to EHR-based measurement; • to have broader applicability (if 1 measure cannot accommodate all patient populations; settings, e.g., hospital, home health; or levels of analysis, e.g., clinician, facility; etc.); • to increase availability of performance results (if 1 measure cannot be widely implemented, e.g., if measures based on different data types increase the number of entities for whom performance results are available) **Note:** Until clinical data from electronic health records (EHRs) are widely available for performance measurement, endorsement of competing measures based on different data types (e.g., claims and EHRs) may be needed to achieve the dual goals of 1) advocating widespread access to performance data and 2) migrating to performance measures based on EHRs. EHRs are the preferred source for clinical record data, but measures based on paper charts or data submitted to registries may be needed in the transition to EHR-based measures. Is an additional measure unnecessary? • primarily for unique developer preferences **Identify the burden of endorsing competing measures** Do the different measures affect interpretability across measures? Does having more than 1 endorsed measure increase the burden of data collection? **Determine if the added value of endorsing competing measures offsets any burden or negative impact** • If yes, recommend competing measures for endorsement (if harmonized) and provide the rationale for recommending endorsement of multiple competing measures. Also, identify analyses needed to conduct a rigorous evaluation of the use and usefulness of the measures at the time of endorsement maintenance. • If no, recommend the best measure for endorsement and provide rationale.

Table 11. Sample Considerations to Justify Lack of Measure Harmonization

Related Measures	Lack of Harmonization	Assess Justification for Conceptual Differences	Assess Justification for Technical Differences
Same measure focus (numerator); different target population (denominator)	Inconsistent measure focus (numerator)	The evidence for the measure focus is different for the different target populations so that 1 measure cannot accommodate both target populations. Evidence should always guide measure specifications.	• Differences in the available data drive differences in the technical specifications for the measure focus. • Effort has been made to reconcile the differences across measures but important differences remain.
Same target population (denominator); different measure focus (numerator)	Inconsistent target population (denominator) and/or exclusions	The evidence for the different measure focus necessitates a change in the target population and/or exclusions. Evidence should always guide measure specifications.	• Differences in the available data drive differences in technical specifications for the target population. • Effort has been made to reconcile the differences across measures but important differences remain.
For any related measures	Inconsistent scoring/ computation	The difference does not affect interpretability or burden of data collection. If it does, it adds value that outweighs any concern regarding interpretability or burden of data collection.	The difference does not affect interpretability or burden of data collection. If it does, it adds value that outweighs any concern regarding interpretability or burden of data collection.

Guidance on Evaluating eMeasures

This guidance is updated from **Review and Update of Guidance for Evaluating Evidence and Measure Testing - Technical Report (October 2013)**

Definition of eMeasure[a] – a measure that is specified in the accepted standard health quality measure format (HQMF) and uses the Quality Data Model (QDM) and value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC). Alternate forms of EHR specifications other than HQMF are not considered eMeasures.[b]

eMeasures are subject to the same evaluation criteria as other performance measures and must meet the criteria that are current at the time of initial submission or endorsement maintenance (regardless of meeting prior criteria or prior endorsement status). Algorithms 1, 2, 3 apply to eMeasures.

A new eMeasure version of an endorsed measure is not considered an endorsed measure until it has been specifically evaluated and endorsed by NQF. An eMeasure should be submitted as a separate measure even if the same or similar measure exists for another data source (e.g., claims or registry). In the near future NQF will update the measure numbering system to include a designation for eMeasures. NQF also plans to link measures that share the same concept except for data source.

Requirements for Endorsing eMeasures

The following guidance addresses and updates the criteria for endorsement of eMeasures.

Specifications

- HQMF specifications are required. The recently released update to HQMF (Release 2 or R2) allows for more complex eMeasures to be specified. Output from the Measure Authoring Tool (MAT) ensures that the measure is in the proper HQMF format; however, the MAT is not required to produce HQMF.
- Value sets.
    - All eMeasures submitted to NQF must have published value sets within the VSAC as part of the measure.
    - If an eMeasure does not have a published value set, then the measure developer must look to see if there is a published value set that aligns with the proposed value set within its measure.
    - If such a published value set does not exist, then the measure developer must demonstrate that the value set is in draft form and is awaiting publication to VSAC.

Each submitted eMeasure undergoes a technical review by NQF staff before going to the Standing Committee for evaluation that includes assessing the measure logic and its conformance to HQMF; a determination if the value sets are in alignment with the scope and purpose of the measure; and that the measure has enough testing data from electronic health records to adequately determine variance in quality among providers.

[a] eMeasures are also known as eCQMs or electronic clinical quality measures.

[b] NQF accepts measures that use EHRs as a data source and that are tested in EHRs (abstraction or local programming) but are not specified and tested with HQMF specifications. These measures, without HQMF specifications, are not considered eMeasures and will be evaluated as traditional measures against the NQF criteria.

Feasibility Assessment

- A feasibility assessment as described in the eMeasure Feasibility Assessment Report (2012) is required for all eMeasures. The feasibility assessment addresses the data elements as well as the measure logic.

Testing for Reliability and Validity

To be considered for NQF endorsement, all eMeasures must be tested for reliability and validity using the HQMF specifications.

- The minimum requirement is testing in **EHR systems from more than 1 EHR vendor**. Developers should test on the number of EHRs they feel appropriate. It is highly desirable that measures are tested in systems from multiple vendors.
- In the description of the sample used for testing, indicate how the eMeasure specifications were used to obtain the data.
- eMeasures specified according to HQMF Release 1 (R1) that have been previously endorsed do not need to be retested. An expansion of the metadata and logic would not fundamentally alter the measure to the point at which retesting is needed. Those measures developed after December 2012, in which the QDM data elements were updated and became the basis for HQMF R2 (in 2013), should be tested in the latest format. eMeasures developed and approved after 2013 will be examined to determine if retesting in HQMF R2 is needed.
- If testing of eMeasures occurs in a small number of sites, it may be best accomplished by focusing on patient-level data element validity (comparing data used in the measure to the authoritative source). However, as with other measures, testing at the level of the performance measure score is encouraged if data can be obtained from enough measured entities. The use of EHRs and the potential access to robust clinical data provide 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 analyses such as sensitivity and specificity, positive predictive value, and negative predictive value. The guidance on measure testing allows this type of validity testing to also satisfy the requirement for reliability testing (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 1 final overall computation for all patients). At a minimum the numerator, denominator, and exclusions (or exceptions) must be assessed and reported separately.
  - Use of a simulated data set is no longer suggested for testing validity of data elements and is best suited for checking that the measure specifications and logic are working as intended.
  - NQF's guidance has some flexibility; therefore, measure developers should consult with NQF staff if they think they have another reasonable approach to testing reliability and validity.
- The general guidance on samples for testing any measure also is relevant for eMeasures:
  - Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.
  - The sample should represent the variety of entities whose performance will be measured. The Measure Testing Task Force recognized that the samples used for reliability and validity testing often

have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.
    - The sample should include adequate numbers of units of measurement *and* adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.
    - When possible, units of measurement and patients within units should be randomly selected.
- The following subcriteria under Scientific Acceptability of Measure Properties also apply to eMeasures.
    - Exclusion analysis (2b3). If exclusions (or exceptions) are not based on the clinical evidence, analyses should identify the overall frequency of occurrence of the exclusions as well as variability across the measured entities to demonstrate the need to specify exclusions.
    - Risk adjustment (2b4). Outcome and resource use measures require testing of the risk adjustment approach.
    - Differences in performance (2b5). This criterion is about using the measure as specified to distinguish differences in performance across the entities that are being measured. The performance measure scores should be computed for all accountable entities for which eMeasure data are available (not just those on which reliability/validity testing was conducted) and then analyzed to identify differences in performance.
    - Because eMeasures are submitted as a separate measures, even if the same or similar measure exists for another data source (e.g., claims), comparability of performance measure scores if specified for multiple data sources (2b6) does not apply.
    - 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.

eMeasure Approval for Trial Use

Developers have indicated that it can be challenging to test eMeasures to the extent necessary to meet NQF endorsement criteria—at least until they have been more widely implemented. At the same time, there is interest in developing eMeasures for use in federal programs and obtaining NQF endorsement for those 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.

In 2014, NQF piloted ***eMeasure Approval for Trial Use*** for eMeasures that are ready for implementation but cannot yet be adequately tested to meet NQF endorsement criteria. NQF uses the multistakeholder 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 for 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 clinical data in EHRs.

In April 2015, the Consensus Standards Approval Committee (CSAC) agreed to make approval for trial use available for all eMeasures submitted to NQF. Approval for trial use is NOT time-limited endorsement as it carries no endorsement label. Measures approved for trial use will be so indicated on QPS. See Table 12 for comparison of endorsement and approval for trial use.

Criteria for approval for trial use include:

- Must meet all criteria under Importance to Measure and Report (clinical evidence and opportunity for improvement/performance gap).
- The eMeasure 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.
- While trial measures are not intended for accountability purposes, there should be a plan for future use and discussion of how the measures will be useful for accountability and improvement.
- Related and competing measures are identified with a plan for harmonization or justification for developing a competing measure.

Maintenance of Trial eMeasures

1. The trial eMeasure designation automatically expires 3 years after initial approval if the eMeasure is not submitted for endorsement prior to that time.
    - The time to submit for endorsement is driven by success with testing. There is no expectation that every trial measure will be submitted for endorsement—some may fail during testing.
    - When submitted for endorsement, the measure will be evaluated through the multistakeholder process. Ideally, Standing Committees and/or more flexible schedules for submitting measures will prevent delays for the endorsement process.
2. If submitted for endorsement prior to the 3-year expiration, the developer can select from the following options for evaluation and endorsement:
    - Option 1: Submit and evaluate only Scientific Acceptability of Measure Properties, including the final eMeasure specifications and all testing. If endorsed, endorsement maintenance will be scheduled from the date approved as a trial measure, at which time it will be submitted for endorsement maintenance and subject to evaluation on all criteria.
    - Option 2: Submit and evaluate on all criteria. If endorsed, a new endorsement date will be identified and endorsement maintenance will be scheduled from the new endorsement date, at which time it will be submitted for endorsement maintenance and subject to evaluation on all criteria
3. If submitted for endorsement 3 or more years after the date of approval as a trial measure, the measure must be submitted and evaluated on all criteria just as any measure being submitted for initial endorsement.

Table 12. Endorsement Versus eMeasure Trial Approval

	Endorsement	eMeasure Trial Approval
Meaning	The eMeasure has been judged to meet all NQF evaluation criteria and is suitable for use in accountability applications as well as performance improvement.	The eMeasure has been judged to meet the criteria that indicate its readiness for implementation in real-world settings in order to generate the data required to assess reliability and validity. Such measures would not have been judged to meet all the criteria indicating it is suitable for use in accountability applications.
Measure Evaluation	Reliability and validity testing results are required upon submission. All criteria are voted on by the Committee. Measure information forms for all measures under review for endorsement are made available on the project webpage.	Reliability and validity testing results are not needed for submission. All other criteria are voted on by the Committee. Measure information forms for all eMeasures under review for trial approval are made available on the project webpage.
Public and Member Comment	Same process. Comments may be submitted on measures recommended and not recommended for endorsement.	Same process. Comments may be submitted on eMeasures recommended and not recommended for eMeasure Trial Approval.
Member Voting	Same process. Members may vote on measures recommended for endorsement.	Same process. Members may vote on measures recommended for eMeasure Trial Approval.
CSAC and BoD	Same process.	Same process.
Information in QPS	Specs for endorsed measures are available.	Specs for eMeasures recommended for trial approval are available

	Endorsement	eMeasure Trial Approval
Status	When due for maintenance review, the measure will be evaluated through the multistakeholder process.	Trial Approval designation expires 3 years after initial approval. When submitted for endorsement, the measure will require testing results and will be evaluated through the multistakeholder process. There are 2 options if submitted for endorsement prior to 3 year expiration: Option 1: Submit and evaluate only Scientific Acceptability of Measure Properties, including the final eMeasure specifications and all testing. If endorsed, endorsement maintenance will be scheduled from the date approved as a trial measure, at which time it will be submitted for endorsement maintenance and subject to evaluation on all criteria. Option 2: Submit and evaluate on all criteria. If endorsed, a new endorsement date will be identified and endorsement maintenance will be scheduled from the new endorsement date, at which time it will be submitted for endorsement maintenance and subject to evaluation on all criteria.

Risk Adjustment for Sociodemographic Factors (SDS) Trial Period

Guidance for Measure Developers

Background Information on the Trial Period

- The NQF Board of Directors approved a 2-year trial period for risk adjustment for sociodemographic factors prior to a permanent change in NQF policy.
- During the trial period, the NQF policy that restricted use of SDS factors in statistical risk models has been suspended, and NQF is implementing several of the **Risk Adjustment Expert Panel's recommendations**.

Instructions for Providing Required Information During the NQF SDS Trial Period

- Enter patient-level sociodemographic variables that were available and analyzed during measure development in Section 1.8 of the Measure Testing Attachment. These variables could include:
    - Patient-reported data (e.g., income, education, language)
    - Proxy variables when sociodemographic data are not collected from each patient (e.g., based on patient address and use of census tract data to assign individual patients to a category of income, education, etc.) and conceptual rationale for use
    - Patient community characteristics (e.g., crime rate, percent vacant housing, smoking rate, level of uninsurance) assigned to individual patients for the specific community where they live (not in the community in which the healthcare unit is located)
- Enter the conceptual description (logical rationale or theory informed by literature and content experts) of the causal pathway between the patient sociodemographic factors, patient clinical factors, quality of care, and outcome in Section 2b4.3 of the Measure Testing Attachment
- Enter the analyses and interpretation resulting in decision to include or not include SDS factors in section 2b4.4b of the Measure Testing Attachment. This analysis could include:
    - Variation in prevalence of the factor across measured entities
    - Empirical association with the outcome (univariate)
    - Contribution of unique variation in the outcome in a multivariable model
    - Assessment of between-unit effects versus within-unit effects to evaluate potential clustering of disadvantaged patients in lower quality units
- Enter reliability and validity testing for the measure as specified in Section 2a2 and 2b2 of the Measure Testing Attachment.
    - If changing from a non-SDS-adjusted risk adjustment model to one that is SDS-adjusted, then *updated* reliability and validity testing is required and must be entered into section 2a2 and 2b2 of the Measure Testing Attachment.
- Enter a comparison of performance scores with and without SDS factors in the risk adjustment model in Section 2b6 of the Measure Testing Attachment.
    - In Section 2b6.1,enter the method of testing conducted to compare performance scores with and without SDS factors in the risk adjustment model for the same entities. Describe the steps and the statistical approach used.
    - In Section 2b6.2, enter the statistical results from testing the differences in the performance scores with and without SDS factors in the risk adjustment model. (e.g., correlation, rank order)
    - In Section 2b6.3, provide an interpretation of your results in terms of the differences in performance scores with and without SDS factors in the risk adjustment model for the same entities. What do the results mean and what are the norms for the test conducted?

    - NOTE: If the measure has more than 1 set of specifications/instructions (e.g., 1 for medical record abstraction and 1 for claims data), then section 2b6 must *also* be used to demonstrate comparability of the performance scores.
- If a performance measure includes SDS variables in its risk adjustment model, the measure developer must provide the information required to stratify a clinically-adjusted-only version of the measure results for those SDS variables in section S.12 in the Measure Submission Form. This information should include *the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate.*
- Enter the details of the final statistical risk model and variables in Sections S.14 and S.15 of the Measure Submission Form.

Guidance on Evaluating Patient-Reported Outcome Performance measures (PRO-PMs)

See NQF report **Patient-Reported Outcomes in Performance Measurement (December 2012)**

Table 13. Distinctions Among PRO, PROM, and PRO-PM: Two Examples

Definition	Patients with Clinical Depression	Persons with Intellectual or Developmental Disabilities
Patient-reported outcome (PRO): The concept of any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. PRO domains encompass: • health-related quality of life (including functional status); • symptom and symptom burden; • experience with care; and • health behaviors.	Symptom: depression	Functional Status-Role: employment
PRO measure (PROM): Instrument, scale, or single-item measure used to assess the PRO concept as perceived by the patient, obtained by directly asking the patient to self-report (e.g., PHQ-9).	PHQ-9©, a standardized *tool* to assess depression	Single-item measure on National Core Indicators Consumer Survey: *Do you have a job in the community?*
PRO-based performance measure (PRO-PM): A performance measure that is based on PROM data aggregated for an accountable healthcare entity (e.g., percentage of patients in an accountable care organization whose depression score as measured by the PHQ-9 improved).	Percentage of patients with diagnosis of major depression or dysthymia and initial PHQ-9 score >9 with a follow-up PHQ-9 score <5 at 6 months (NQF #0711)	The proportion of people with intellectual or developmental disabilities who have a job in the community

Table 14. NQF Endorsement Criteria and their Application to PRO-PMs

Abbreviated NQF Endorsement Criteria	Considerations for Evaluating PRO-PMS That Are Relevant to Other Performance Measures	Unique Considerations for Evaluating PRO-PMS
1. Importance to Measure and Report a. Evidence: Health outcome OR evidence-based intermediate outcome, process, or structure of care b. Performance gap c. High priority d. Composite	• PRO-PMs should have the same evidence requirement as health outcomes—rationale supports the relationship of the health outcome to processes or structures of care. • Exceptions to the evidence requirement for performance measures focused solely on administering a PROM should be addressed the same as other measures based solely on conducting an assessment (e.g., order lab test, check BP).	• Patients/persons must be involved in identifying PROs for performance measurement (person-centered; meaningful).
2. Scientific Acceptability of Measure Properties a. Reliability 1. Precise specifications 2. Reliability testing (data elements or performance measure score) b. Validity 1. Specifications consistent with evidence 2. Validity testing (data elements or performance measure score) 3. Exclusions 4. Risk adjustment 5. Identify differences in performance 6. Comparability of multiple sets of specifications 7. Missing data/non-response	• Data collection instruments (tools) should be identified (e.g., specific PROM instrument, scale, or single item). • If multiple data sources (i.e., PROMs, methods, modes, languages) are used, then comparability or equivalency of performance measure scores should be demonstrated.	• Specifications should include standard methods, modes, languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; how missing data are handled; and calculation of response rates to be reported with the performance measure results. • Reliability and validity should be demonstrated for <u>both</u> the data (PROM) and the PRO-PM performance measure score. • Differences in individuals' PROM values related to PROM instruments or methods, modes, and languages of administration need to be analyzed and potentially included in risk adjustment. • Response rates can affect validity and should be addressed in testing.
3. Feasibility a. Data generated and used in care delivery b. Electronic data c. Data collection strategy can be implemented	• The burdens of data collection, including those related to use of proprietary PROMs, are minimized and do not outweigh the benefit of performance measurement.	• The burden to respondents (people providing the PROM data) should be minimized (e.g., availability and accessibility enhanced by multiple languages, methods, modes). • Infrastructure to collect PROM data and integrate into workflow and EHRs, as appropriate.
4. Usability and Use a. Accountability and transparency b. Improvement c. Benefits outweigh unintended negative consequences	• Adequate demonstration of the criteria specified above supports usability and ultimately the use of a PRO-PM for accountability and performance improvement.	

Abbreviated NQF Endorsement Criteria	Considerations for Evaluating PRO-PMS That Are Relevant to Other Performance Measures	Unique Considerations for Evaluating PRO-PMS
5. Comparison to Related or Competing Measures **5a.** Harmonization of related measures **5b.** Competing measures	• Apply to PRO-PMs	• PRO-PMs specified to use different PROM instruments will be considered competing measures

Guidance on Evaluating Composite Performance Measures

Definition

A composite performance measure is a combination of 2 or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

Box 1. Identification of Composite Performance Measures for Purposes of NQF Measure Submission, Evaluation, and Endorsement*

The following **will be** considered composite performance measures for purposes of NQF endorsement:

- Measures with 2 or more individual performance measure scores combined into 1 score for an accountable entity.
- Measures with 2 or more individual component measures ***assessed separately for each patient*** and then aggregated into 1 score for an accountable entity. These include:
    - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient); or
    - any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient).

The following **will not be** considered composite performance measures for purposes of NQF endorsement at this time:

- Single performance measures, even if the data are patient scores from a composite instrument or scale (e.g., single performance measure on communication with doctors, computed as the percentage of patients where the average score for 4 survey questions about communication with doctors is equal to or greater than 3).
- Measures with multiple measure components that are assessed for each patient, but that result in multiple scores for an accountable entity, rather than a single score. These generally should be submitted as separate measures and indicated as paired/grouped measures.
- Measures of multiple linked steps in 1 care process assessed for each patient. These measures focus on 1 care process (e.g., influenza immunization) but may include multiple steps (e.g., assess immunization status, counsel patient, and administer vaccination). These are distinguished from all-or-none composites that capture multiple care processes or outcomes (e.g., foot care, eye care, glucose control).
- Performance measures of 1 concept (e.g., mortality) specified with a statistical method or adjustment **(e.g., empirical Bayes shrinkage estimation)** that combines information from the accountable entity with information on average performance of all entities or a specified group of entities (e.g., by case volume), **typically in order to increase reliability**.

* The list in Box 1 includes the types of measure construction most commonly referred to as composites, but this list is not exhaustive. NQF staff will review any potential composites that do not clearly fit 1 of these descriptions and make the determination of whether the measure will be evaluated against the additional criteria for composite performance measures.

Table 15. NQF Measure Evaluation Criteria and Guidance for Evaluating Composite Performance Measures

Abbreviated NQF Endorsement Criteria	Guidance For Composite Performance Measures
1. Importance to Measure and Report a. Evidence: Health outcome OR evidence-based intermediate outcome, process, or structure of care b. Performance gap c. For composite performance measures, the following must be explicitly articulated and logical: 1. 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 2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and 3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.	The evidence subcriterion (1a) must be met for each component of the composite (unless NQF-endorsed under the current evidence requirements). The evidence could be for a group of interventions included in a composite performance measure (e.g., studies in which multiple interventions are delivered to all subjects and the effect on the outcomes is attributed to the group of interventions). The performance gap criterion (1b) must be met for the composite performance measure as a whole. The performance gap for each component also should be demonstrated. However, if a component measure has little opportunity for improvement, justification for why it should be included in the composite is required (e.g., increase reliability of the composite, clinical evidence). 1c. Must also be met for a composite performance measure to meet the must-pass criterion of Importance to Measure and Report. If the developer provides a conceptual justification as to why an "any-or-none" measure should not be considered a composite, and that justification is accepted by the NQF steering committee, the measure can then be considered a single measure rather than a composite.
2. Scientific Acceptability of Measure Properties a. Reliability 1. Precise specifications 2. Reliability testing (data elements or performance measure score) b. Validity 1. Specifications consistent with evidence 2. Validity testing (data elements or performance measure score) 3. Exclusions 4. Risk adjustment 5. Identify differences in performance 6. Comparability of multiple sets of specifications 7. Missing data/non-response 2c. Disparities 2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that: 1. 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 2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible; and 3. the extent of missing data and how the specified handling of missing data minimizes bias (i.e., achieves scores that are an accurate	Composite measure specifications include component measure specifications (unless individually endorsed); scoring rules (i.e., how the component scores are combined or aggregated); how missing data are handled (if applicable); required sample sizes (if applicable); and when appropriate, methods for standardizing scales across component scores and weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite). **2a2.** For composite performance measures, reliability must be demonstrated for the composite measure score. Testing should demonstrate that measurement error is acceptable relative to the quality signal. Examples of testing include signal-to-noise analysis, interunit reliability, and intraclass correlation coefficient. Demonstration of the reliability of the individual component measures is not sufficient. In some cases, component measures that are not independently reliable can contribute to reliability of the composite measure. **2b2.** For composite performance measures, validity should be empirically demonstrated for the composite measure score. If empirical testing is not feasible at the time of initial endorsement, acceptable alternatives include systematic assessment of content or face validity of the composite performance measure or demonstration that each of the component measures meet NQF subcriteria for validity. By the time of endorsement maintenance, validity of the composite performance measure must be empirically demonstrated. It is unlikely that a "gold standard" criterion exists, so validity testing generally will focus on construct validation—testing hypotheses based

Abbreviated NQF Endorsement Criteria	Guidance For Composite Performance Measures
reflection of quality).	on the theory of the construct. Examples include testing the correlation with measures hypothesized to be related or not related; testing the difference in scores between groups known to differ on quality assessed by some other measure. **2b3.** Applies to the component measures and composite performance measures. **2b4.** Applies to outcome component measures (unless NQF-endorsed). **2b5.** Applies to composite performance measures. **2b6.** Applies to component measures. **2b7.** Analyses of overall frequency of missing data and distribution across providers. Ideally, sensitivity analysis of the effect of various rules for handling missing data and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules. **2c.** Applies to composite performance measures. **2d.** Must also be met for a composite performance measure to meet the must-pass criterion of Scientific Acceptability of Measure Properties. If empirical analyses do not provide adequate results (or are not conducted), other justification must be provided and accepted for the measure to potentially meet the must-pass criterion of Scientific Acceptability of Measure Properties. **Examples** of analyses: 1. ***If components are correlated*** – analyses based on shared variance (e.g., factor analysis, Cronbach's alpha, item-total correlation, mean inter-item correlation). 1. ***If components are not correlated*** – analyses demonstrating the contribution of each component to the composite score (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable, or clinical justification (e.g., correlation of the individual component measures to a common outcome measure). 2. Ideally, sensitivity analyses of the effect of various considered aggregation and weighting rules and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.
3. Feasibility a. Data generated and used in care delivery b. Electronic data c. Data collection strategy can be implemented	**3a, 3b, 3c.** Apply to composite performance measures as a whole, taking into account all component measures.

Abbreviated NQF Endorsement Criteria	Guidance For Composite Performance Measures
4. Usability and Use a. Accountability and transparency b. Improvement c. Benefits outweigh unintended negative consequences	Note that NQF endorsement applies only to the composite performance measure as a whole, not to the individual component measures (unless they are submitted and evaluated for individual endorsement). **4a.** Applies to composite performance measures. To facilitate transparency, at a minimum, the individual component measures of the composite must be listed with use of the composite measure. **4b.** Applies to composite performance measures. **4c.** Applies to composite performance measures and component measures. If there is evidence of unintended negative consequences for any of the components, the developer should explain how that is handled or justify why that component should remain in the composite.
5. Comparison to Related or Competing Measures **5a.** Harmonization of related measures **5b.** Competing measures	**5a and 5b.** Apply to composite performance measures as a whole as well as the component measures.

Guidance for Evaluating Evidence for Measures of Appropriate Use

Measures for appropriate use of procedures and medical technologies are becoming more common and reflect multistakeholder interest in assessing appropriate use of healthcare services. Current NQF criteria and guidance regarding appropriate use measures indicate the following:

- NQF measure evaluation criteria state that evidence for measures that focus on inappropriate use should include "a systematic assessment and grading of the quality, quantity, and consistency of the body of evidence that the measured process *does not* lead to a desired health outcome." Thus, the evidence for appropriate/inappropriate use measures should primarily focus on the *lack of effectiveness or benefit* of the test or procedure to patients. Patient safety considerations such as unnecessary exposure to radiation or anesthesia or complications from inappropriate tests or procedures may contribute to the risk-benefit evidence.
- Cost and resource use are **not** the focus of appropriate use measures. The cost and resource use implications of appropriate use measures are no different than for other measures; for example, improvement in adverse outcomes after surgery will likely reduce costs; and improved use of screening tests will increase costs but this is not a consideration for evaluating the measures.
- Appropriate use measures are not efficiency measures as currently defined by NQF (i.e., efficiency measures per the current NQF definition have both a quality component and a cost component in the measure construct).

<u>Development of Appropriate Use Method</u>

In the 1980's, RAND/UCLA developed a methodology to determine "appropriateness" of healthcare tests, procedures, and processes. This method has been used worldwide in a variety of medical applications and

forms the basis of many appropriate use measures (AUM) submitted to NQF. The RAND/UCLA Appropriateness Method User's Manual (2001) defines

> "an appropriate procedure as one in which "the expected health benefit (e.g., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeds the expected negative consequences (e.g., mortality, morbidity, anxiety, pain, time lost from work) by a sufficiently wide margin that the procedure is worth doing, exclusive of cost." The rationale behind the method is that randomized clinical trials—the "gold standard" for evidence-based medicine—often either are not available or cannot provide evidence at a level of detail sufficient to apply to the wide range of patients seen in everyday clinical practice. Although robust scientific evidence about the benefits of many procedures is lacking, physicians must nonetheless make decisions every day about when to apply them. Consequently, the RAND/UCLA researchers believed a method was needed that would combine the best available scientific evidence with the collective judgment of experts to yield a statement regarding the appropriateness of performing a procedure at the level of patient-specific symptoms, medical history and test results."

Various specialty societies such as the American College of Radiology and the American College of Cardiology Foundation/American Heart Association have used the RAND/UCLA methodology to develop appropriate use criteria for imaging and cardiovascular technology. The American Academy of Orthopedic Surgeons and the American Academy of Dermatology have also established appropriate use criteria for aspects of their specialty. These specialty society guidelines are intended to guide clinicians in the appropriate use of various tests and procedures.

Clinical Practice Guidelines and Appropriate Use Criteria

The appropriate use criteria are guidelines for clinical practice. The method for developing appropriate use criteria is very similar to the method used to develop traditional clinical practice guidelines (CPGs). Table 16 presents a side-by-side comparison of the methods for developing CPGs and Appropriate Use Criteria (AUC). Development of both types of guidelines is based on a review of the evidence.

Table 16. Comparison of Development of CPGs and AUCs

Clinical Practice Guidelines	Appropriate Use Criteria
Generally disease- or condition-based	Generally procedure- or test-based
Methodology: Institute of Medicine "Clinical Practice Guidelines We Can Trust" "The processes by which a CPG is developed and funded should be detailed explicitly and publicly accessible."	Methodology: RAND/UCLA Appropriateness Method (RAM)

Clinical Practice Guidelines	Appropriate Use Criteria
Evidence review: CPG developers should use systematic reviews that meet standards set by the IOM's Committee on Standards for Systematic Reviews of Comparative Effectiveness Research: • A summary of relevant available evidence (and evidentiary gaps), description of the quality (including applicability), quantity (including completeness), and consistency of the aggregate available evidence. • A clear description of potential benefits and harms. • A rating of the level of confidence in (certainty regarding) the evidence underpinning the recommendation.	Evidence review: • Fundamental to any appropriateness study is a critical review of the literature summarizing the scientific evidence available on the procedure under review. Literature reviews for appropriateness studies are typically less strict in their inclusion criteria, as the objective is to produce a synthesis of all the information available on a particular topic; where evidence from controlled trials is lacking, they may well include lower-quality evidence from, for example, cohort studies or case series. • Where possible, "evidence tables" summarizing the data from multiple studies should be included in the literature review.
Guideline development group (GDG) composition: • The GDG should be multidisciplinary and balanced, comprising a variety of methodological experts and clinicians, and populations expected to be affected by the CPG. • Whenever possible GDG members should not have conflicts of interest. • Funders should have no role in CPG development.	Expert panel: • Most users of the RAND/UCLA method recommend using multidisciplinary panels to better reflect the variety of specialties that are actually involved in patient treatment decisions. • The RAM is a modified Delphi method that, unlike the original Delphi, provides panelists with the opportunity to discuss their judgments between the rating rounds.

Clinical Practice Guidelines	Appropriate Use Criteria
Guideline Recommendations: • Recommendations should include an explanation of the reasoning underlying the recommendation. • A rating of the strength of the recommendation in light of the evidence. • A description and explanation of any differences of opinion regarding the recommendation. • Recommendations should be articulated in a standardized form detailing precisely what the recommended action is and under what circumstances it should be performed. • Strong recommendations should be worded so that compliance with the recommendation(s) can be evaluated. • The CPG publication date, date of pertinent systematic evidence review, and proposed date for future CPG review should be documented in the CPG.	RAND/UCLA Appropriateness Method (RAM): • A list of the hypothetical clinical scenarios or "indications" to be rated by the panel is developed. The purpose of the list of indications is to classify patients in terms of the clinical variables physicians take into account in deciding whether to recommend a particular procedure. • Panelists are asked to rate the appropriateness of each indication using their own best clinical judgment (rather than their perceptions of what other experts might say) and considering an average patient presenting to an average physician who performs the procedure in an average hospital (or other care-providing facility). *They are specifically instructed not to consider cost implications in making their judgments.* Although cost considerations are an important factor in deciding whether a procedure or treatment should ultimately be made available to patients, the RAM focuses on the initial question of whether it is effective. • In the RAM a procedure is classified as "appropriate," "uncertain," or "inappropriate" for a particular patient scenario ("indication") in accordance with 1) the *median* panel rating and 2) some measure of the dispersion of panel ratings, which is taken as an indicator of the level of agreement with which the ratings were made. [This is not a consensus process.]

NQF's Evaluation Criteria for Evidence

NQF's guidance for evidence for measures in general, and specifically those based on clinical practice guidelines, applies to measures based on appropriateness criteria as well. As noted in Table 16 above, both CPGs and appropriateness methodologies require systematic reviews of the evidence generated from a thorough literature search.

Measure Submission

Measure submitters should provide the information on evidence that was provided to the expert panel that developed the appropriate use criteria, along with any updated evidence published since the AUC was developed. The measure submission should include:

- a summary (not a list of references) of the evidence in the submission evidence attachment that describes the quantity, quality, and consistency of the body of evidence (not selected references) and an assessment of the benefits versus harms; and

- a link to (or an attached appendix that contains) the complete evidence report with evidence tables, if available.

Committee Evaluation

Committees should review the information provided and evaluate the evidence presented according to Algorithm 1.

- It is unlikely that a systematic review will have been performed to establish a lack of benefit for an intervention. Begin at Box 7 – empiric evidence submitted without systematic review and grading of the evidence.
- If a complete literature review is summarized (rather than selected studies – Box 8) then the Committee should decide whether the submitted evidence indicates a **high certainty** and that benefits clearly outweigh undesirable effects (Box 9). If yes, then rate as moderate.
- If there is no empiric evidence, skip Box 10 and go to Box 11. The Committee should agree that the AUC method is a systematic assessment of expert opinion that the benefits of what is being measured outweigh the potential harms (Box 11). If the Committee agrees that it is acceptable (or beneficial) to hold providers accountable for performance in the absence of empiric evidence (Box 12), then rate as "insufficient evidence with exception."

Inactive Endorsement with Reserve Status (November 2014)

Given the number of publicly reported measures with high levels of performance, reliable and valid measures of great importance may not retain NQF endorsement due to the lack of a performance gap. The purpose of an inactive endorsement with reserve status is to retain endorsement of reliable and valid quality performance measures that have overall high levels of performance with little variability so that performance could be monitored as necessary to ensure that performance does not decline. This status would apply only to highly credible, reliable, and valid measures that have high levels of performance due to incorporation into standardized patient care processes and quality improvement actions. The key issue for continued endorsement is the opportunity cost associated with continued measurement at high levels of performance—rather than focusing on areas with known gaps in care. Endorsement with reserve status retains these measures in the NQF portfolio for periodic monitoring, while also communicating to potential users that the measures no longer address high-leverage areas for accountability purposes.

Measures with High Levels of Performance — Recommendations from the Evidence Task Force

The 2010 Evidence Task Force defined the term "topped out" as meaning that there are high levels of performance with little variation and, therefore, little room for further improvement. The Task Force did not recommend specific quantitative thresholds for identifying conformance with the subcriterion opportunity for improvement (1b). Threshold values for opportunity for improvement would be difficult to standardize and depend on the size of the population at risk, the effectiveness of an intervention, and the consequences of the quality problem**.** For example, even modest variation would be sufficient justification for some highly effective, potentially life-saving treatments (e.g., certain vaccinations) that are critical to the public health.

The Task Force noted that, at the time of endorsement maintenance review, if measure performance data indicate overall high performance with little variation, then justification would be required for continued

endorsement of the measure. The Consensus Standards Approval Committee (CSAC) added that the default action should be to remove endorsement unless there is a strong justification to continue endorsement. If a measure fails opportunity for improvement (1b), then it does not pass the threshold criterion, *Importance to Measure and Report,* and is therefore not suitable for endorsement.

Task Force recommendations related to opportunity for improvement (1b) include the following:

- At the time of initial endorsement, evidence for opportunity for improvement generally will be based on research studies, or on epidemiologic or resource use data. However, at the time of review for endorsement maintenance, the primary interest is on the endorsed measure as specified, and the *evidence for opportunity for improvement should be based on data for the specific endorsed measure.*
- When assessing measure performance data for opportunity for improvement, the following factors should be considered:
    - number and representativeness of the entities included in the measure performance data;
    - data on disparities; and
    - size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.
- In exceptional situations, a strong justification for continued endorsement could be considered (e.g., ***evidence*** that overall performance will likely deteriorate if not monitored, magnitude of potential harm if outcomes deteriorate while not being monitored).

Criteria for Assigning Inactive Endorsement with Reserve Status to Measures with High Levels of Performance

There is rarely evidence that performance will deteriorate if a measure is not monitored; therefore, some additional criteria are needed. The following criteria are to be used when there are concerns that performance will deteriorate, but no evidence. These criteria are intentionally rigorous so that the use of endorsement with reserve status is by exception.

- Evidence of little opportunity for improvement (1b), i.e., overall high level of performance with little variation. When assessing measure performance data for opportunity for improvement, the following factors should be considered:
    - distribution of performance scores;
    - number and representativeness of the entities included in the measure performance data;
    - data on disparities; and
    - size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.
- Evidence for measure focus (1a) – there should be strong direct evidence of a link to a desired health outcome; therefore, there would be detrimental consequence on patient health outcomes if performance eroded. Generally, measures more distal to the desired outcome have only indirect evidence of influence on the outcome and would not qualify for reserve endorsement status.
  For process and structure measures, the measure focus should be proximal to the desired outcome. Generally, measures more distal to the desired outcome would not be eligible for reserve status.
- Reliability (2a) – high or moderate rating: Reliability has been demonstrated for the measure score.
- Validity (2b) – high or moderate: Validity has been demonstrated by empiric testing for the measure score (face validity not acceptable).
- The reason for high levels of performance is better performance, not an issue with measure construction/specifications (e.g., “documentation”).

- Demonstrated usefulness for improving quality (e.g., data on trends of improvement and scope of patients and providers included).
- Demonstrated use of the measure (e.g., specific programs and scope of patients and providers included); would not grant inactive endorsement status for a measure that has not been used).
- If a measure is found to be "topped out," i.e., does not meet criteria for opportunity for improvement (1b), the measure will only be considered for inactive endorsement with reserve status. The measure must meet all other criteria as noted above, otherwise the measure should not be endorsed.

Maintenance of Inactive Endorsement with Reserve Status

Measures assigned inactive endorsement status will not be reviewed in the usual endorsement maintenance review cycle. During portfolio review the Standing Committee will periodically review measures in reserve status for any change in evidence, evidence of deterioration in performance or unintended consequences, or any other concerns related to the measure. The Standing Committee may remove a measure from inactive endorsement status if the measure no longer meets NQF endorsement criteria. A maintenance review may occur upon a request from the Standing Committee or measure steward to return the measure to active endorsement.

Measures in reserve status will be considered for harmonization with related or competing measures. Measure developers should be aware of measures in reserve status and avoid developing duplicative measures.