NATIONAL QUALITY FORUM

Measure Evaluation Criteria

Effective July 2013 (updated 10/11/13)

The criteria include the approved recommendations from the 2012 projects on patient-reported outcomes, composites, and eMeasure feasibility.

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 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 three 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. A measure that has not been tested for reliability and validity is only potentially eligible for time-limited endorsement if all of the following conditions are met: 1) the measure topic is not addressed by an endorsed measure; 2) it is relevant to a critical timeline (e.g., legislative mandate) for implementing endorsed measures; 3) the measure is not complex (requiring risk adjustment or a composite); and 4) the measure steward verifies that testing will be completed within 12 months of endorsement.

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, Performance Gap, and Priority (Impact)—Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows: Guidance-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 not required for the resource use component.

AND

1b. Performance Gap
Demonstration of quality problems and opportunity for improvement, i.e., data [7] demonstrating

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

AND

1c. High Priority (previously referred to as High Impact)
The measure addresses:

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

OR

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

meaningful.

AND

1d. For composite performance measures, the following must be explicitly articulated and logical:
1d1. The quality construct, including the overall area of quality; included component measures; and the relationship of the component measures to the overall composite and to each other; and
1d2. The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually; and
1d3. How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

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

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***.

2a. Reliability
2a1. The measure is well defined and precisely specified [8] so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and must use the Quality Data Model (QDM) and value sets vetted through the National Library of Medicine's Value Set Authority Center (VASC). [9]

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

2b. Validity
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 (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

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

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

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

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

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

Notes
8. Measure specifications include the target population (denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (numerator, target condition, event, outcome), measurement time window, exclusions, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, scoring/computation. Specifications for **PRO-PMs** also include: specific PROM(s); standard methods, modes, and languages of administration; whether (and

how) proxy responses are allowed; standard sampling procedures; handling of missing data; and calculation of response rates to be reported with the performance measure results.
Specifications for **composite performance measures** include: component measure specifications (unless individually endorsed); aggregation and weighting rules; handling of missing data; standardizing scales across component measures; required sample sizes.
9. If HQMF or the QDM does not support all aspects of a particular measure construct (e.g., risk adjustment, composite aggregation and weighting rules), those aspects may be specified outside HQMF with an explanation and plans to request expansion of the relevant standards. If a value set is not vetted by the VSAC, explain why and plans to submit for approval. eMeasure specifications include data type from the QDM, value sets and attributes, measure logic, original source of the data and recorder.
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

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.

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

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.

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

Notes
17. All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with

measures based on patient surveys and when there are small numbers of patients.
18. The feasibility assessment uses a standard score card or a fully transparent alternative that includes at a minimum: a description of the assessment; feasibility scores for all data elements on availability, accuracy, standard coding system, and workflow; explanatory notes for all data element components scoring a "1" (lowest rating); measure logic can be executed; and rationale and plan for addressing feasibility concerns.

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.

4a. Accountability and Transparency
Performance results are used in at least one accountability application [1] within three years after initial endorsement and are publicly reported [19] within six years after initial endorsement (or the data on performance results are available). [20] If not in use at the time of initial endorsement, then a credible plan [21] for implementation within the specified timeframes is provided.

AND

4b. Improvement
Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. [22] 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.

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).

Notes
19. 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.
20. 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.
21. 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.
22. 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.

5. Comparison to Related or Competing Measures
If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both 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.