NATIONAL QUALITY FORUM

Measure Evaluation Criteria
December 2009

Conditions for Consideration
Four conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards:
A. The measure is in the public domain or an intellectual property agreement is signed.
B. The measure owner/steward verifies there is an identified responsible entity and 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 public reporting and quality improvement.
D. The requested measure submission information is complete. Generally, measures should be fully developed and tested so that all the evaluation criteria have been addressed and information needed to evaluate the measure is provided. Measures that have not been tested are only potentially eligible for a time-limited endorsement and in that case, measure owners must verify that testing will be completed within 12 months of endorsement.

Criteria for Evaluation
If all four conditions for consideration are met, candidate measures are evaluated for their suitability based on four sets of standardized criteria: importance to measure and report, scientific acceptability of measure properties, usability, and feasibility. Not all acceptable measures will be strong—or equally strong—among each set of criteria. The assessment of each criterion is a matter of degree; however, all measures must be judged to have met the first criterion, importance to measure and report, in order to be evaluated against the remaining criteria.

1. Importance to measure and report: Extent to which the specific measure focus is important to making significant gains in health care quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes for a specific high impact aspect of healthcare where there is variation in or overall poor performance. ***Candidate measures must be judged to be important to measure and report*** *in order to be evaluated against the remaining criteria.*

1a. The measure focus addresses:
- a specific national health goal/priority identified by NQF's National Priorities Partners;
  OR
- a demonstrated high impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, high resource use (current and/or future), severity of illness, and patient/societal consequences of poor quality).

1b. Demonstration of quality problems and opportunity for improvement, i.e., data[1] demonstrating considerable variation, or overall poor performance, in the quality of care across providers and/or population groups (disparities in care).

1c. The measure focus is:
- an outcome (e.g., morbidity, mortality, function, health-related quality of life) that is relevant to, or

[1] Examples of data on opportunity for improvement include, but are not limited to: prior studies, epidemiologic data, measure data from pilot testing or implementation. If data are not available, the measure focus is systematically assessed (e.g., expert panel rating) and judged to be a quality problem.

associated with, a national health goal/priority, the condition, population, and/or care being addressed[2];
OR

- if an intermediate outcome, process, structure, etc., there is **evidence**[3] that supports the specific measure focus as follows:
  - o Intermediate outcome – evidence that the measured intermediate outcome (e.g., blood pressure, Hba1c) leads to improved health/avoidance of harm or cost/benefit.
  - o Process – evidence that the measured clinical or administrative process leads to improved health/avoidance of harm and
    if the measure focus is on one step in a multi-step care process[4], it measures the step that has the greatest effect on improving the specified desired outcome(s).
  - o Structure – evidence that the measured structure supports the consistent delivery of effective processes or access that lead to improved health/avoidance of harm or cost/benefit.
  - o Patient experience – evidence that an association exists between the measure of patient experience of health care and the outcomes, values and preferences of individuals/ the public.
  - o Access – evidence that an association exists between access to a health service and the outcomes of, or experience with, care.
  - o Efficiency[5] – demonstration of an association between the measured resource use and level of performance with respect to one or more of the other five IOM aims of quality.

If not important to measure and report, STOP.

2. Scientific acceptability of the measure properties: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

2a. The measure is well defined and precisely specified[6] so that it can be implemented consistently within and across organizations and allow for comparability. The required data elements are of high quality as defined by NQF's Health Information Technology Expert Panel (HITEP) [7] .

---

[2] Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, "never events" that are compared to zero are appropriate outcomes for public reporting and quality improvement.

[3] The strength of the body of evidence for the specific measure focus should be systematically assessed and rated (e.g., USPSTF grading system – grade definitions and methods). If the USPSTF grading system was not used, the grading system is explained including how it relates to the USPSTF grades or why it does not. However, evidence is not limited to quantitative studies and the best type of evidence depends upon the question being studied (e.g., randomized controlled trials appropriate for studying drug efficacy are not well suited for complex system changes). When qualitative studies are used, appropriate qualitative research criteria are used to judge the strength of the evidence.

[4] 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 multi-step process, the step with the greatest effect on the desired outcome should be selected as the focus of measurement. For example, although assessment of immunization status and recommending immunization are necessary steps, they are not sufficient to achieve the desired impact on health status – patients must be vaccinated to achieve immunity. This does not preclude consideration of measures of preventive screening interventions where there is a strong link with desired outcomes (e.g., mammography) or measures for multiple care processes that affect a single outcome.

[5] Efficiency of care is a measurement construct of cost of care or resource utilization associated with a specified level of quality of care. It is a measure of the relationship of the cost of care associated with a specific level of performance measured with respect to the other five IOM aims of quality. Efficiency might be thought of as a ratio, with quality as the numerator and cost as the denominator. As such, efficiency is directly proportional to quality, and inversely proportional to cost. (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; based on AQA Principles of Efficiency Measures).

[6] Measure specifications include the target population (e.g., denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (e.g., numerator),

2b. Reliability testing[8] demonstrates the measure results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period.

2c. Validity testing[9] demonstrates that the measure reflects the quality of care provided, adequately distinguishing good and poor quality. If face validity is the only validity addressed, it is systematically assessed.

2d. Clinically necessary measure exclusions are identified and must be:

- supported by evidence[10] of sufficient frequency of occurrence so that results are distorted without the exclusion;

AND

- a clinically appropriate exception (e.g., contraindication) to eligibility for the measure focus[11];

AND

- precisely defined and specified:
  - if there is substantial variability in exclusions across providers, the measure is specified so that exclusions are computable and the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion);
  - if patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that it strongly impacts performance on the measure and the measure must be specified so that the information about patient preference and the effect on the measure is transparent[12] (e.g., numerator category computed separately, denominator exclusion category computed separately).

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

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified and is based on patient clinical factors that influence the measured outcome (but not disparities in care) and are present at start of care[11,13]

---

measurement time window, exclusions, risk adjustment, definitions, data elements, data source and instructions, sampling, scoring/computation.

[7] The HITEP criteria for high quality data include: a) data captured from an authoritative/accurate source; b) data are coded using recognized data standards; c) method of capturing data electronically fits the workflow of the authoritative source; d) data are available in EHRs; and e) data are auditable. NQF. *Health Information Technology Expert Panel Report: Recommended Common Data Types and Prioritized Performance Measures for Electronic Healthcare Information Systems*. Washington, DC: NQF; 2008.

[8] Examples of reliability testing 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 may address the data items or final measure score.

[9] Examples of validity testing include, but are not limited to: determining if measure scores adequately distinguish between providers known to have good or poor quality assessed by another valid method; correlation of measure scores with another valid indicator of quality for the specific topic; ability of measure scores to predict scores on some other related valid measure; content validity for multi-item scales/tests. Face validity is a subjective assessment by experts of whether the measure reflects the quality of care (e.g., whether the proportion of patients with BP < 140/90 is a marker of quality). If face validity is the only validity addressed, it is systematically assessed (e.g., ratings by relevant stakeholders) and the measure is judged to represent quality care for the specific topic and that the measure focus is the most important aspect of quality for the specific topic.

[10] Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers.

[11] Risk factors that influence outcomes should not be specified as exclusions.

[12] Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

[13] 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, gender (e.g., poorer treatment outcomes of

OR
- rationale/data support no risk adjustment.

2f. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[14] differences in performance.

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

2h. 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);
OR
rationale/data justifies why stratification is not necessary or not feasible.

3. Usability: Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

3a. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audience(s) for both public reporting (e.g., focus group, cognitive testing) and informing quality improvement (e.g., quality improvement initiatives)[15]. 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.

3b. The measure specifications are harmonized[16] with other measures, and are applicable to multiple levels and settings.

3c. Review of existing endorsed measures and measure sets demonstrates that the measure provides a distinctive or additive value to existing NQF-endorsed measures (e.g., provides a more complete picture of quality for a particular condition or aspect of healthcare).

4. Feasibility: Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

4a. For clinical measures, required data elements are routinely generated concurrent with and as a

---

African American men with prostate cancer, inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than adjusting out differences.

[14] 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% v. 75%) 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 poor performance may not demonstrate much variability across providers.

[15] Public reporting and quality improvement are not limited to provider-level measures – community and population measures also are relevant for reporting and improvement.

[16] Measure harmonization refers to the standardization of specifications for similar measures on the same topic (e.g., ***influenza immunization*** of patients in hospitals or nursing homes), or related measures for 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 dictated by the evidence. The dimensions of harmonization can include numerator, denominator, exclusions, 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.

byproduct of care processes during care delivery.

4b. The required data elements are available in electronic sources. If the required data are not in existing electronic sources, a credible, near-term path to electronic collection by most providers is specified and clinical data elements are specified for transition to the electronic health record.

4c. Exclusions should not require additional data sources beyond what is required for scoring the measure (e.g., numerator and denominator) unless justified as supporting measure validity.

4d. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified.

4e. Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality[17], etc.) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use).

If a measure meets the above criteria <u>and</u> there are competing measures (either endorsed measures, or other new submissions that also meet the criteria), compare measures on: Scientific acceptability of measure properties, Usability, and Feasibility to determine best-in-class.

5. Demonstration that the measure is superior to competing measures – new submissions and/or endorsed measures (e.g., is a more valid or efficient way to measure).

[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.