NATIONAL QUALITY FORUM

Measure Evaluation Criteria and Additional Guidance for Population Health Measures
January 2012

NQF Measure Evaluation Criteria *(updated January 2011)*	***Population Health Measure Evaluation: Additional Guidance and Context****
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.**	**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.	**A.** No change.
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.	**B.** The measure owner/steward verifies there is an identified responsible entity or multi-stakeholder entities and a process to maintain and update the measure on a schedule that is commensurate with the rate of population health innovation, but at least every three years.
C. The intended use of the measure includes both public reporting and quality improvement.	**C.** The intended use of the measure includes both public reporting and improvement in efforts to improve population health.
D. The measure is fully specified and tested for reliability and validity.[1]	D. No change**.**
E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate.	E. The measure developer/steward attests that harmonization with related measures and issues with competing measures have been considered and addressed, as appropriate. **Harmonization of related measures at the provider and population levels measures has been considered and addressed.**
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.	F. No change.

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

Note **1.** 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, candidate measures are evaluated for their suitability based on four sets of standardized criteria in the following order: *Importance to Measure and Report*, *Scientific Acceptability of Measure Properties*, *Usability*, and *Feasibility*. Not all acceptable measures will be equally strong among 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.	**Criteria for Evaluation** No change.
1. Impact, Opportunity, Evidence—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-impact aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to*** *meet all three subcriteria to pass this criterion and be evaluated against the remaining criteria.* **1a. High Impact** The measure focus addresses:	**1. Impact, Opportunity, Evidence—Importance to Measure and Report:** Extent to which the specific measure focus is evidence-based, important to making significant gains in **population health, improving determinants of health and health outcomes of a population** for a high-impact aspect of **health** where there is variation in **(including geographic variation)** or overall less-than-optimal performance. *Measures must be judged to meet all three subcriteria to pass this criterion and be evaluated against the remaining criteria.* **1a. High Impact** **Note: For population health measures, high impact**

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

• a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; **OR** • a demonstrated 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). **AND** **1b. Performance Gap** Demonstration of quality problems and opportunity for improvement, i.e., data[2] demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers and/or population groups (disparities in care). **AND** **1c. Evidence to Support the Measure Focus** The measure focus is a health outcome or is evidence-based, demonstrated as follows: • Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care. • Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading	**would also be identified by the National Prevention Strategy and the DHHS Consensus Statement on Quality in Public Health.** **OR** • a demonstrated high-impact aspect of **health** (e.g., affects **large population** and/or has a substantial impact for a smaller population; **source of significant health disparities**; leading cause of morbidity/mortality; functional health; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality). **AND** **1b. Performance Gap** Demonstration of opportunity for improvement in **health**, i.e., data[2] demonstrating considerable variation, or overall less-than-optimal performance, in **health across providers (healthcare, public health, and other partners) and/or population groups, (including but not limited to disparities in care.** **AND** **1c. Evidence to Support the Measure Focus** • Health outcome:[3] a rationale supports the relationship of **the health outcomes in the population to strategies to improve health.** • **Health determinant**, Intermediate outcome, Process, or Structure: a systematic assessment and grading of the quantity, quality,

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome. • Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes. • Efficiency:[6] evidence for the quality component as noted above. **Notes** **2.** 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. **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.** 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. **5.** 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,	and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome. • **Experience with care, services or other health determinants**: evidence that the measured aspects of care are those valued by **people and populations** and for which the **respondent** is the best and/or only source of information OR that experience is correlated with desired outcomes. • Efficiency:[6] evidence for the quality component as noted above. **4. Population health determinants typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with stakeholder input) → provide intervention → evaluate impact on population health status. If the measure focus is one step in such a multistep process, the steps with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement.** **5.** No change.

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

Development and Evaluation (GRADE) guidelines. **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).	**6.** No change.
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[7] so it can be implemented consistently within and across organizations and allow for comparability. EHR measure specifications are based on the quality data model (QDM).[8] **2a2.** Reliability testing[9] 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. **2b. Validity** **2b1.** The measure specifications[7] are consistent with the evidence presented to support the focus of measurement under criterion 1c. 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[10] 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.	**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[7] so it can be implemented consistently within and across organizations, **multistakeholder groups, populations or entities with shared accountability for health** and allow for comparability. **2a2.** No change. **2b. Validity.** **2b1.** No change. **2b2.** Validity testing[10] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the effect of interventions to **improve population health**, adequately identifying differences in effectiveness.

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

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;[11] **AND** If patient preference (e.g., informed decision-making) 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).[12]	**2b3.** Exclusions are supported by the **evidence**; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; **AND** If **individual or subgroup** preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure **or variation**; in such cases, the measure must be specified so that the information about **individual or subgroup** preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).
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 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;[13,14] and has demonstrated adequate discrimination and calibration **OR** • rationale/data support no risk adjustment/ stratification.	**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 factors that influence the measured outcome (but not factors related to **disparities in population health or health interventions**) and are present at start of care;[13,14] and has demonstrated adequate discrimination and calibration
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[15] differences in performance; **OR** there is evidence of overall less-than-optimal performance.	**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 meaningful[15] differences in performance **or variation across populations in improving health.** **OR** there is evidence of overall less-than-optimal performance **or significant variation across populations**.

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.	**2b6.** No change.
2c. Disparities 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);	**2c. Disparities** If **health disparities** 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	**OR**
rationale/data justifies why stratification is not necessary or not feasible.	**No option for justification for lack of stratification.**
Notes	**Notes**
7. 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.	**7.** No change
8. EHR measure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.	**8.** N/A
9. 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).	**9.** No change.
10. Validity testing applies to both the data	**10.** No change.

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

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.	
11. 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.	**11.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, **multistakeholder groups, and populations** and sensitivity analyses with and without the exclusion.
12. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.	**12.** N/A
13. Risk factors that influence outcomes should not be specified as exclusions.	**13.** Risk factors that influence outcomes should not be specified as exclusions.
14. 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	**14.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in health determinants, such as race, socioeconomic status, or gender (e.g., poorer **health outcomes** of African American men with prostate cancer or inequalities in CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

out the differences. **15.** 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.	differences. **15.** With large enough sample sizes, small differences that are statistically significant may or may not be **practically meaningful**. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of **people** who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is meaningful; or whether a statistically significant difference of $25 in cost for an intervention (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers **or populations**.
3. Usability: Extent to which intended audiences (e.g., consumers, purchasers, providers, policymakers) can understand the results of the measure and find them useful for decision-making. **3a**. Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for public reporting (e.g., focus group, cognitive testing) or rationale; **AND** **3b.** Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for informing quality improvement[16] (e.g., quality improvement initiatives) or rationale. **Note** **16.** 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	**3. Usability: Note: intended audiences can include community members and coalitions.** **3a.** Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for public reporting (e.g., focus group, cognitive testing) or rationale; **AND** **3b.** Demonstration that information produced by the measure is meaningful, understandable, and useful to the intended audiences for informing improvement[16] in **health determinants and/or population health** or rationale. **Note** **16.** An important outcome that may not have an identified improvement strategy still can be useful for informing **improvement in quality and/or population health** by identifying the need for and

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

improvement.	stimulating new approaches to improvement.
4. Feasibility: Extent to which the required data are readily available or could be captured without undue burden and can be implemented for performance measurement. **4a.** For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order). **4b.** 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. **4c.** Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified. **4d.** 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). **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.	**4. Feasibility:** No change. **4a.** No change for clinically oriented measures. **4b.** The required data elements are available in electronic health records, **personal health records, health information exchanges, population data bases**, or other electronic sources. If the required data are not available **in existing electronic sources**, a credible, near-term path to electronic collection is specified. 4c. Susceptibility to inaccuracies, errors, **inappropriate comparison across populations,** or unintended consequences and the ability to audit the data items to detect such problems are identified. **4d.** No change. **Note** **17.** All data collection must conform to laws regarding protected health information. **Confidentiality** is of particular concern with measures based on **individual** surveys and **for small populations**.
5. Comparison to Related or Competing Measures	**5. Comparison to Related or Competing Measures**

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**

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[18] 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** **18.** 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.	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. **Note: Complementary measures that address different improvement strategies are not considered competing measures.** **OR** **5b.** No change. **Note** **18. Additional conceptualization needed for harmonization between clinical and population-level measures.**

*** While NQF's measure evaluation criteria (left column) can be extended to population-level measurement, additional guidance and context are required to address conceptual and methodological issues specific to population-level performance measurement. This information is included in the right column, in red font.**