Resource Use Measure Evaluation Criteria (Version 1.2)*

BACKGROUND

The resource use measure evaluation criteria are grounded in the standard NQF evaluation criteria, keeping the four major criteria (importance, scientific acceptability, usability, and feasibility) in place but modifying the subcriteria as appropriate to reflect the specific needs of resource use measure evaluation. The notes for the subcriteria have also been updated to provide specific guidance around meeting the criteria for resource use measures, including appropriate data analysis methods and clarification of concepts.

Resource Use Measure Evaluation Criteria
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 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 <u>both</u> public reporting <u>and</u> quality improvement. D. The measure is fully specified and tested for reliability and validity. *Based on existing NQF policy, complex measures are not eligible for time-limited endorsement. Resource use measures are complex and therefore must be fully tested at the time of submission.* 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.
Criteria for Evaluation If all conditions for consideration are met, candidate consensus standards 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 the 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.

This criteria is subject to change pending changes to the NQF Measure Evaluation Criteria for quality measures.

Resource Use Measure Evaluation Criteria (Version 1.2)*

Resource Use Measure Evaluation Criteria
1. Importance to measure and report
Resource use measures will be evaluated based on the extent to which the specific measure focus is important to making significant contributions toward understanding healthcare costs for a specific high-impact aspect of healthcare where there is variation or a demonstrated high-impact aspect of healthcare (e.g., affects large numbers, leading cause of morbidity/mortality, variation in resource use [current and/or future], severity of illness, and patient/societal consequences of poor quality) or overall poor performance. Candidate consensus standards 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 DHHS or the National Priorities Partnership convened by NQF: OR – a demonstrated high-impact aspect of healthcare[1] (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). AND 1b. Demonstration of resource use or cost problems and opportunity for improvement, i.e., data[1] demonstrating variation[2] in the delivery of care across providers and/or population groups (disparities in care). AND 1c. The purpose/objective of the resource use measure[3] (including its components) and the construct for resource use/costs are clearly described. AND 1d. The resource use service categories (i.e., types of resources/costs) that are included in the resource use measure are consistent with and representative of the conceptual construct represented by the measure. Whether or not the resource use measure development begins with a conceptual construct or a set of resource service categories, the service categories included must be conceptually coherent and consistent with the purpose.

Resource Use Measure Evaluation Criteria
2. Scientific acceptability of the measure properties
Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the cost or resources used to deliver care.
2a. Reliability 2a1. The measure is well defined and precisely specified[4] so that it can be implemented consistently within and across organizations and allow for comparability. Electronic health record (EHR) measure specifications are based on the quality data set (QDS).[5] 2a2. Reliability testing[6] demonstrates that 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, and/or that the measure score is precise. 2b. Validity 2b1. The measure specifications[4] are consistent with the evidence presented to support the focus of measurement under criterion *1b*. 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[7] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the cost of care or resources provided, adequately distinguishing higher and lower cost or resource use. 2b3. Exclusions are supported by the clinical evidence[8] otherwise, they are supported by evidence[9] of sufficient frequency of occurrence so that results are distorted with the exclusion; AND – Measure specifications for scoring include computing exclusions so that the effect on the measure is transparent (i.e., impact clearly delineated, such as number of cases excluded, exclusion rates by type of exclusion); 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[10] (e.g., numerator category computed separately, denominator exclusion category computed separately). 2b4. 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 factors related to disparities in care or the quality of care) and are present at start of care[11,12] and has demonstrated adequate discrimination and calibration OR – rationale/data support no risk-adjustment/-stratification.

2b5. Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful[13] differences in performance,
OR
there is evidence of overall less than optimal performance.

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

2c. 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.

Resource Use Measure Evaluation Criteria
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. The measure performance results are reported to the public at large in national or community reporting programs by the time of endorsement maintenance review. Exceptions may be considered if there is evidence that the measure performance results are available for public reporting and that use of the measure has benefited the public. AND 3b. The measure performance results are considered meaningful, understandable, and useful to the intended audience(s) for both public reporting and informing quality improvement (supported by rationale or demonstration). 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. 3c. Data and result detail are maintained such that the resource use measure, including the clinical and construction logic for a defined unit of measurement can be decomposed to facilitate transparency and understanding. 3d. The measure specifications are harmonized[14] with related measures, OR the differences in specifications are evaluated to be justified.

Resource Use Measure Evaluation Criteria
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 related to measurement are judged to be inconsequential or can be minimized through proper actions, OR can be monitored and detected. 4d. The data collection and measurement strategy can be implemented as demonstrated by operational use in external reporting programs, OR testing did not identify barriers to operational use (e.g., barriers related to data availability, timing, frequency, sampling, patient confidentiality,[15] fees for use of proprietary specifications).

NOTES FOR RESOURCE USE MEASURE EVALUATION CRITERIA

Notes for Importance

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 or performance problem.
2. Findings from peer-reviewed literature and empirical data are examples of acceptable information that can be used to justify importance and demonstrating variation. It is the proof of the measure's concept that enables the Committee to determine if the measure is valid in addressing this concept.
3. Resource use measures are broadly applicable and comparable measures of input counts (in terms of units or dollars) applied to a population or population sample. Resource use measures count the frequency of specific resources; these resource units may be monetized as appropriate. The approach to monetizing resources varies and often depends on the perspective of the measurer and those being measured. Monetizing resource use allows for the aggregation across resources. Related to resource use, 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 multidimensional concept that includes inputs and outputs, and specifically the amount of resources used (the inputs) and the degree of quality achieved (output)—resource use measures alone do not capture efficiency but are a building block of efficiency: Efficiency = fx (quality, resource use). 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. 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. (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; based on the AQA Principles of Efficiency Measures (http://www.aqaalliance.org/files/PrinciplesofEfficiencyMeasurementApril2006.doc).

Notes for Scientific Acceptability

4. Well-defined, complete, and precise specifications for resource use measures include three of the specification modules: measure clinical logic and method, measure construction logic, and adjustments for comparability as relevant to the measure. Data protocol steps are critical to the reliability and validity of the measure; specifications must be detailed enough such that users can execute the necessary steps to implement the measure. Further, additional sub-functions within the data protocol and measure reporting modules may require precise specifity as indicated on the submission form and as appropriate to the

submitted measure. To allow for flexibility of measure implementation, clear guidance from the measure developer is required at time of measure submission on those data protocol and measure reporting steps that are not specified with the measure; this guidance will be reviewed for adequacy by the review Committees. For those modules and analytic functions that are required in the submission form that the measure developer deems as not relevant or available, justification for and implications of not specifying those steps is required. Specifications should also include the identification of the target population to whom the measure applies, identification of those from the target population who achieved the specific measure focus (i.e., target condition, event), measurement time window, exclusions, risk-adjustment, definitions, data elements, data source and instructions, sampling, scoring/computation. The resource use measure submission form is the platform through which this information is submitted.

5. EHR measure specifications include data type from the QDS, code lists, EHR field, measure logic, original source of the data, recorder, and setting.
6. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; and test-retest, split-half reliability. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise). Refer to the 2010 Testing Task Force report, which provides detailed guidance for measure testing and evaluating the scientific acceptability of measure properties. Reliability for resource use measures can be demonstrated for each of the modules or for the entire measure score. For those modules not included in the reliability testing, justification for and implications of not addressing those steps is required.
7. Validity testing applies to both the data elements and the 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 measure's scores indicate resource use, e.g., measure scores are different for groups known to have differences in resource use assessed by another valid resource use measure or method; correlation of measure scores with another valid indicator of resource use for the specific topic; or relationship to conceptually related 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 higher from lower resource use or costs. Refer to the 2010 Testing Task Force report, which provides detailed guidance for measure testing and evaluating the scientific acceptability of measure properties. The scoring/aggregation and weighting rules used during measure scoring and construction are consistent with the conceptual construct. If differential weighting is used, it should be justified. Differential weights are determined by empirical analyses or a

systematic assessment of expert opinion or value-based priorities. Validity testing for resource use measures can be used to demonstrate validity for each module or the entire measure score. For those modules not included in the demonstration of validity, justification for and implications of not addressing those steps is required.

8. Examples of evidence that exclusion distorts measure results include, but are not limited to: frequency or cost of occurrence, sensitivity analyses with and without the exclusion, and variability of exclusions across providers. Some measures may specify the exclusion of some patients, events, or episodes that are known or determined to be high cost. For example, a patient with active cancer may be excluded from a chronic obstructive pulmonary disease (COPD) resource use measure because cancer is considered to be the dominant medical condition with known high costs. Or an episode that exceeds a specified threshold (e.g., 3 standard deviations from the mean) relative to episodes of the same type may be excluded with the recommendation for those high-cost episodes to be examined separately. Exclusions must be justified and supported with appropriate evidence on the effect of the exclusions. Testing for resource use measure exclusions should address the appropriate specification steps (i.e., clinical logic, and thresholds and outliers). For those exclusions not addressed, justification for and implications of not addressing them is required. Exclusions do not include the algorithms used to identify the population or area of measurement (e.g., if the measure examines diabetes, exclusion testing does not include those patients without diabetes).
9. 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.
10. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions. 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, then 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 (e.g., numerator category computed separately, denominator exclusion category computed separately). Patient co-pays or co-amounts should not exclude a service from inclusion, or justification to exclude these patients or services should be provided. Specifically, claims for services received by the patient should be included for episode construction or patient identification and resource use or cost assessment even when the patient pays a portion of the claims, unless otherwise justified—all approaches should be transparent.
11. Risk factors that influence outcomes or resource use/cost should not be specified as exclusions; exclusions for resource use or cost that influence results must be justified.

12. 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 African American men with prostate cancer, inequalities in treatment for cardiovascular disease risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out differences.
13. 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.

Notes for Usability

14. 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), or 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, 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.

Notes for Feasibility

15. 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.