NATIONAL QUALITY FORUM

Measure Evaluation Criteria and Guidance Summary Tables
Effective for Projects Beginning after January 2011

Contents

Process for Evaluating and Recommending Measures

Measures considered as potential voluntary consensus standards are evaluated by a multistakeholder steering committee against four major criteria (*Importance to Measure and Report, Scientific Acceptability of Measure Properties, Usability*, and *Feasibility*). Each criterion has several subcriteria that are used to determine if the criterion is met and all the subcriteria must be evaluated. The evaluation criteria, subcriteria, explanatory footnotes, and evaluation guidance should be thoroughly reviewed prior to evaluating measures.

Measure stewards/developers submit measures for consideration in a standardized form that is structured to solicit the information necessary for committees to determine whether the NQF criteria are met. The submission form is their opportunity to demonstrate that that the criteria are met.

Committee members first review and evaluate the measures individually, but ultimately the entire Steering Committee as a group determines to what extent the criteria are met and whether to recommend measures for endorsement by NQF. NQF recognizes that each committee member brings different expertise and experience to the project and may not feel qualified to evaluate all aspects of a measure. All committee members should contribute to the evaluation to the best of their ability, knowing that the final evaluation rating and recommendation will be made by the full Steering Committee.

Preliminary Evaluation by Individual Committee Members

Depending on the number of measures to be evaluated, committee members may be assigned all or a subset of the measures for review and in-depth preliminary evaluation prior to the meeting of the entire committee. However, all committee members will participate in the evaluation of all measures.

All assigned measures should be evaluated on all criteria and subcriteria prior to the meeting and entered into the online tool. The rating scale and definitions are provided after each criterion in the preceding information. Committee members are encouraged to review and evaluate as many of the other measures as possible prior to the meeting and may also enter any additional evaluations into the online tool. For some projects, a technical advisory panel (TAP) advises the Steering Committee on the extent to which the subcriteria are met; the TAP evaluates only the subcriteria and its evaluations are provided to the steering committee.

All the committee members' preliminary evaluations will be compiled and distributed for use by the committee at the meeting, so discussions can be focused on questions and potential areas of disagreement. ***Assigned reviewers will be asked to begin the discussion of a few assigned measures by briefly describing the measure, summarizing the compiled preliminary evaluation ratings and rationales, highlighting areas of concern and differences of opinion***.

Evaluation by the Entire Committee

At the in-person committee meeting, measure stewards are asked to briefly introduce their group of measures and also are available to respond to questions raised by the Steering Committee; however, the discussion, evaluation, and recommendation of measures are the purview of the committee alone.

Each measure will be introduced by a committee member as described above then discussed by the entire committee. After the committee's discussion, the entire steering committee will vote on the rating for each of the four major criteria and finally, on whether the measure meets the NQF criteria for

endorsement. Related and competing measures are addressed only if measures are considered suitable for endorsement.

Occasionally, committee members may identify a modification thought to be necessary to make a measure suitable for endorsement. Because the Committee does not develop measures and changes in measure specifications may have implications for the testing results, recommended changes in tested measures are not routine. In such cases, the committee will first vote on the measure as it was submitted; if it does not pass, then it can be voted on with a condition that must be addressed by the steward/developer before further consideration.

If a measure was voted as meeting the NQF criteria for endorsement and there are no measure harmonization issues or competing measures, the measure is recommended for endorsement. Otherwise, harmonization of related measures and selection of the best measure from among competing measures must be addressed before a final recommendation is made. Measures which are only eligible for time-limited endorsement will be identified by NQF prior to voting.

The Steering Committee's recommendations are followed by a draft report that is posted for NQF member and public comment, NQF member voting, review and approval by the Consensus Standards Approval Committee (CSAC), Board endorsement, and opportunity for appeals.

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 public reporting and quality improvement.

D. The measure is fully specified and tested for reliability and validity.[1]

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. 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.

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.* Yes☐ No☐ Guidance-Table 3

1a. High Impact H☐ M☐ L☐ I ☐ Definitions-Table 5

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 (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 H☐ M☐ L☐ I ☐ Definitions-Table 5

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 Quantity: Yes☐ No☐ Guidance-Table 3

Quantity: H☐ M☐ L☐ I ☐ Quality: H☐ M☐ L☐ I ☐ Consistency: H☐ M☐ L☐ I ☐ Guidance-Table 2

The measure focus is a health outcome or 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.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading 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, 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).

Guidance on Evaluating Importance to Measure and Report

For more information, see: Guidance for Evaluating the Evidence Related to the Focus of Quality Measurement and Importance to Measure and Report

Table 1: Evidence to Support the Focus of Measurement

Type of Measure	Evidence	Example of Measure Type and Evidence to Be Addressed
Health Outcome An outcome of care is the health status of a patient (or change in health status) resulting from healthcare— desirable or adverse. In some situations, resource use may be considered a proxy for a health state (e.g., hospitalization may represent deterioration in health status).	A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service. See Table 5.	**#0230** Acute myocardial Infarction 30-day mortality Survival is a goal of seeking and providing treatment for AMI. **Rationale** linking healthcare processes/ interventions (aspirin, reperfusion) to mortality/ survival **#0171** Acute care hospitalization (risk-adjusted) [of home care patients] Improvement or stabilization of condition to remain at home is a goal of seeking and providing home care services. **Rationale** linking healthcare processes (e.g., medication reconciliation, care coordination) to hospitalization of patients receiving home care services **#0140** Ventilator-associated pneumonia for ICU and high-risk nursery (HRN) patients Avoiding harm from treatment is a goal when seeking and providing healthcare. **Rationale** linking healthcare processes (e.g., ventilator bundle) to ventilator acquired pneumonia
Intermediate Clinical Outcome An intermediate outcome is a change in physiologic state that leads to a longer-term health outcome.	Quantity, quality, and consistency of a body of evidence that the measured intermediate clinical outcome leads to a desired health outcome. See Table 4.	**#0059** Hemoglobin A1c management [A1c > 9] **Evidence** that hemoglobin A1c level leads to health outcomes (e.g., prevention of renal disease, heart disease, amputation, mortality)
Process A process of care is a healthcare-related activity performed for, on behalf of, or by a patient.	Quantity, quality, and consistency of a body of evidence that the measured healthcare process leads to desired health outcomes in the target population with benefits that outweigh harms to patients. Specific drugs and devices should have FDA approval for the target condition. If the measure focus is on inappropriate use, then quantity, quality, and consistency	**#0551** ACE inhibitor/Angiotensin receptor blocker (ARB) use and persistence among members with coronary artery disease at high risk for coronary events **Evidence** that use of ACE-I and ARB results in lower mortality and/or cardiac events **#0058** Inappropriate antibiotic treatment for adults with acute bronchitis

Type of Measure	Evidence	Example of Measure Type and Evidence to Be Addressed
	of a body of evidence that the measured healthcare process does *not* lead to desired health outcomes in the target population. See Table 4.	**Evidence** that antibiotics are not effective for acute bronchitis
Structure Structure of care is a feature of a healthcare organization or clinician related to its capacity to provide high-quality healthcare.	Quantity, quality, and consistency of a body of evidence that the measured healthcare structure leads to desired health outcomes with benefits that outweigh harms (including evidence for the link to effective care processes and the link from the care processes to desired health outcomes). See Table 4.	**#0190** Nurse staffing hours **Evidence** that higher nursing hours result in lower mortality or morbidity, or leads to provision of effective care processes (e.g., lower medication errors) that lead to better outcomes
Special Considerations by Topic		
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 (often acquired through qualitative studies) OR • Evidence that patient experience with care is correlated with desired outcomes	**#0166** HCAHPS **Evidence** that patients/consumers value the aspects of care being measured (e.g., communication with doctors and nurses, responsiveness of hospital staff, pain control, communication about medicines, cleanliness and quiet of the hospital environment, and discharge information)
Efficiency Measures of efficiency combine the concepts of resource use *and* quality	Efficiency measured with combination of quality measures and resource use measures **Quality measure component:** Evidence for the selected quality measure(s) as described in this table **Resource use measure component:** Does not require clinical evidence as described in this table	Currently, there are no NQF-endorsed efficiency measures that combine quality and resource use. **Potential measure:** Diabetes quality measure(s) or composite used in conjunction with a measure of resource use per episode **Evidence** for diabetes quality measure(s) as described in this table

Table 2: Evaluation of Quantity, Quality, and Consistency of Body of Evidence for Structure, Process, and Intermediate Outcome Measures

Definition/ Rating	**Quantity** of Body of Evidence	**Quality** of Body of Evidence	**Consistency** of Results of Body of Evidence
Definition	Total number of studies (not articles or papers)	Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence related to study factors[a] including: study design or flaws; directness/indirectness to the specific measure (regarding the population, intervention, comparators, outcomes); imprecision (wide confidence intervals due to few patients or events)	Stability in both the direction and magnitude of clinically/practically meaningful benefits and harms to patients (benefit over harms) across studies in the body of evidence
High	5+ studies[b]	Randomized controlled trials (RCTs) providing direct evidence for the specific measure focus, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction and similar in magnitude across the preponderance of studies in the body of evidence
Moderate	2-4 studies[b]	• Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect OR • RCTs without serious flaws that introduce bias, but with either indirect evidence or imprecise estimate of effect	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction across the preponderance of studies in the body of evidence, but may differ in magnitude If only one study, then the estimate of benefits greatly outweighs the estimate of potential harms to patients (one study cannot achieve high consistency rating)
Low	0-1 studies[b]	• RCTs with flaws that introduce bias OR • Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations	• Estimates of clinically/practically meaningful benefits and harms to patients differ in both direction and magnitude across the preponderance of studies in the body of evidence OR • wide confidence intervals prevent estimating net benefit If only one study, then estimate of benefits do not greatly outweigh harms to patients
Insufficient to Evaluate *(See Table 5 for exceptions.)*	• No empirical evidence OR • Only selected studies from a larger body of evidence	• No empirical evidence OR • Only selected studies from a larger body of evidence	No assessment of magnitude and direction of benefits and harms to patients

[a] *Study designs* that affect certainty of confidence in estimates of effect include: randomized controlled trials (RCTs), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders.
Study flaws that may bias estimates of effect include: lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; and failure to report important outcomes.

Imprecision with wide confidence intervals around estimates of effects can occur in studies involving few patients and few events.
Indirectness of evidence includes: indirect comparisons (e.g., two drugs compared to placebos rather than head-to head); and differences between the population, intervention, comparator interventions, and outcome of interest and those included in the relevant studies.[15]
[b]The suggested number of studies for rating levels of quantity is considered a general guideline.

Table 3: Evaluation of Subcriterion 1c Based on the Quantity, Quality, and Consistency of the Body of Evidence

Quantity of Body of Evidence	**Quality** of Body of Evidence	**Consistency** of Results of Body of Evidence	**Pass Subcriterion 1c**
Moderate-High	Moderate-High	Moderate-High	Yes
Low	Moderate-High	Moderate (if only one study, high consistency not possible)	Yes, but only if it is judged that additional research is unlikely to change conclusion that benefits to patients outweigh harms; otherwise, No
Moderate-High	Low	Moderate-High	Yes, but only if it is judged that potential benefits to patients clearly outweigh potential harms; otherwise, No
Low-Moderate-High	Low-Moderate-High	Low	No
Low	Low	Low	No
Exception to Empirical Body of Evidence for Health Outcome For a health outcome measure: A rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service			Yes, if it is judged that the rationale supports the relationship of the health outcome to at least one healthcare structure, process, intervention, or service
Potential Exception to Empirical Body of Evidence for Other Types of Measures If there is no empirical evidence, expert opinion is systematically assessed with agreement that the benefits to patients greatly outweigh potential harms.			Yes, but only if it is judged that potential benefits to patients clearly outweigh potential harms; otherwise, No

Table 4: Evidence for Evaluating Importance to Measure and Report

Pass Criterion, Importance to Measure and Report?			
All three subcriteria (1a, 1b, 1c) must be met to pass the threshold criterion, *Importance to Measure and Report.*			
Subcriterion	**Evidence**	**Example**	**Pass the Subcriterion?**
High impact (1a)	• Addresses a *specific* national health goal/priority identified by the Secretary of DHHS or the NPP OR • Epidemiologic or resource use data; health services research – affects large numbers of patients and/or has a very substantial impact for smaller populations; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and patient/societal consequences of poor quality	#0140 Ventilator-associated pneumonia for ICU and high-risk nursery (HRN) patients **NPP goal:** Focus relentlessly on continually reducing and seeking to eliminate all healthcare-associated infections (HAIs) **Evidence** related to numbers of patients (e.g., 250,205 VAPs reported; 35,969 (14.4%) were fatal; cost (e.g., total annual cost of VAP $2.5 billion)	**Yes—** Demonstrated at least one of the aspects of high impact (**High or moderate rating** described in Table 5) **No—**Did not demonstrate at least one of the aspects of high impact
Opportunity for improvement (1b)	**Initial Endorsement** Epidemiologic or resource use data or health services research demonstrating considerable variation or overall less than optimal performance for the focus of measurement across providers and/or population groups (disparities in care) **Review for Endorsement Maintenance** Data for the measure as specified and endorsed demonstrating considerable variation or overall less than optimal performance	#0432 Influenza vaccination of nursing home/skilled nursing facility residents **NPP goal:** All Americans will receive the most effective preventive services recommended by the U.S. Preventive Services Task Force. **Evidence** that vaccination rates vary (e.g., 39% fail to reach the Healthy People 2010 objective of vaccinating at least 90% of nursing home residents)	**Yes—** Demonstrated either variation or overall less than optimal performance (**High or moderate rating** described in Table 5) **No—**Did not demonstrate either variation **or** overall less than optimal performance
Evidence for the focus of measurement (1c)	See Table 2	See Table 2	See Table 2 and Table 3

Table 5: Generic Scale for Rating Subcriteria 1a and 1b

Rating	**Definition**
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

2. Reliability and Validity—Scientific Acceptability of Measure Properties
Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to** meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.* Yes☐ No☐ Guidance-Table 7

2a. Reliability H☐ M☐ L☐ I☐ Guidance-Table 6; EHR measures-Table 8
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 H☐ M☐ L☐ I☐ Guidance-Table 6; EHR measures-Table 8
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.

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 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).[12]

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.

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.

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

2c. Disparities H☐ M☐ L☐ I☐ Definitions-Table 9
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.

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.
8. EHR measure specifications include data type from the QDM, code lists, EHR field, measure logic, original source of the data, recorder, and setting.
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).
10. 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.
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.
12. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
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 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.

Guidance on Evaluating Scientific Acceptability of Measure Properties

For more information, see full report: Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties

Table 6: Evaluation Ratings for Reliability and Validity

Rating	Reliability	Validity
High	All measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring, etc.) are unambiguous and likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.; **AND** Empirical evidence of reliability of **BOTH** data elements (Table A-2) **AND** measure score (Table A-1) within acceptable norms: • Data element: appropriate method, scope, and reliability statistics for critical data elements within acceptable norms (new testing, or prior evidence for the same data type); **OR** commonly used data elements for which reliability can be assumed (e.g., gender, age, date of admission); **OR** *may forego data element reliability testing if data element validity (Table A-4) was demonstrated*; **AND** • Measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) are consistent with the evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of **BOTH** data elements (Table A-4) **AND** measure score (Table A-3) within acceptable norms: • Data element: appropriate method, scope, and statistical results within acceptable norms (new testing, or prior evidence for the same data type) for critical data elements; **AND** • Measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased
Moderate	All measure specifications are unambiguous as noted above **AND** Empirical evidence of reliability within acceptable norms for either critical data elements **OR** measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either critical data elements **OR** measure score as noted above; **OR** Systematic assessment of face validity of measure score as a quality indicator (as described in Table A-3) explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased
Low	One or more measure specifications (e.g., numerator, denominator, exclusions, risk factors, scoring) are ambiguous with potential for confusion in identifying who is included and excluded from the target	The measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and

	population, or the event, condition, or outcome being measured; or how to compute the score, etc.; **OR** Empirical evidence (using appropriate method and scope) of unreliability for either data elements **OR** measure score, i.e., statistical results outside of acceptable norms	scope) of invalidity for either data elements **OR** measure score, i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias results
Insufficient Evidence	Inappropriate method or scope of reliability testing	Inappropriate method or scope of validity testing (including inadequate assessment of face validity as noted above); **OR** Threats to validity as noted above are likely and are NOT empirically assessed

Table 7: Evaluation of Scientific Acceptability of Measure Properties Based on Reliability and Validity Ratings

Validity Rating	**Reliability Rating**	**Pass *Scientific Acceptability of Measure Properties* for Initial Endorsement***	
High	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	**Moderate-High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Low	Any rating	No	Validity of conclusions about quality is the primary concern. If evidence of validity is rated low, the reliability rating will usually also be low. Low validity and moderate-high reliability represents inconsistent evidence.

*A measure that does not pass the criterion of *Scientific Acceptability of Measure Properties* would not be recommended for endorsement.

Table 8: Evaluation of Reliability and Validity of Measures Specified for EHRs

	New Measure Specified for EHR		Modifications for Endorsed Measure *Re-specified* for EHRs
Rating	**Reliability Description and Evidence**	**Validity Description and Evidence**	
High	All EHR measure specifications are unambiguous[+] and include only data elements from the Quality Data Model (QDM)***** including quality data elements, code lists, and measure logic; **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability of both data element **AND** measure score within acceptable norms: • Data element: reliability (repeatability) assured with computer programming—**must test data element validity** **AND** • Measure score: appropriate method, scope, and reliability statistic within acceptable norms	The measure specifications (numerator, denominator, exclusions, risk factors) reflect the quality of care problem (1a,1b) and evidence cited in support of the measure focus (1c) under *Importance to Measure and Report*; **AND** Empirical evidence of validity of both data elements **AND** measure score within acceptable norms: • Data element: validity demonstrated by analysis of agreement between data elements electronically extracted and data elements visually abstracted from the entire EHR with statistical results within acceptable norms; **OR** complete agreement between data elements and computed measure scores obtained by applying the EHR measure specifications to a simulated test EHR data set with known values for the critical data elements; **AND** • Measure score: appropriate method, scope, and validity testing result within acceptable norms; **AND** Identified threats to validity (lack of risk adjustment/stratification, multiple data types/methods, systematic missing or "incorrect" data) are empirically assessed and adequately addressed so that results are not biased	The EHR measure specifications use only data elements from the Quality Data Model (QDM)***** and include quality data elements, code lists, and measure logic; **AND** Crosswalk of the EHR measure specifications (QDM quality data elements, code lists, and measure logic) to the endorsed measure specifications demonstrates that they represent the original measure, which was judged to be a valid indicator of quality; **AND** Analysis of comparability of scores produced by the retooled EHR measure specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limits
Moder-ate	All EHR measure specifications are unambiguous[+] and include only data elements from the QDM;* **OR** new data elements are submitted for inclusion in the QDM; **AND** Empirical evidence of reliability within acceptable norms for either data elements **OR** measure score as noted above	The measure specifications reflect the evidence cited under *Importance to Measure and Report* as noted above; **AND** Empirical evidence of validity within acceptable norms for either data elements **OR** measure score as noted above; **OR** Systematic assessment of face validity of measure score as a quality indicator (as described in Table A-3) explicitly addressed and found substantial agreement that ***the scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality*** **AND** Identified threats to validity noted above are empirically assessed and adequately addressed so that results are not biased	The EHR measure specifications use only data elements from the QDM as noted above **AND** Crosswalk of the EHR measure specifications as noted above demonstrates that they represent the original measure **AND** For measures with time-limited status, testing of the original measure and evidence ratings of moderate for reliability and validity as described in Table 2.
Low	One or more EHR measure specifications are ambiguous[+] or do not use data elements from the QDM*; **OR** Empirical evidence of unreliability for either data elements **OR** measure score—i.e., statistical results outside of acceptable norms	The EHR measure specifications do not reflect the evidence cited under *Importance to Measure and Report* as noted above; **OR** Empirical evidence (using appropriate method and scope) of invalidity for either data elements **OR** measure score— i.e., statistical results outside of acceptable norms **OR** Identified threats to validity noted above are empirically assessed and determined to bias results	The EHR measure specifications do not use only data elements from the QDM; **OR** Crosswalk of the EHR measure specifications as noted above identifies that they do not represent the original measure **OR** For measures with time-limited status, empirical evidence of low reliability or validity for original time-limited measure
Insuffi	Inappropriate method or scope	Inappropriate method or scope of validity testing	Crosswalk of the EHR measure

	New Measure Specified for EHR		Modifications for Endorsed Measure *Re-specified* for EHRs
Rating	**Reliability Description and Evidence**	**Validity Description and Evidence**	
cient evidence	of reliability testing	(including inadequate assessment of face validity as noted above) **OR** Threats to validity as noted above are likely and are NOT empirically assessed	specifications as noted above was not completed OR For measures with time-limited status, inappropriate method or scope of reliability or validity testing for original time-limited measure

[+]Specifications are considered unambiguous if they are likely to consistently identify who is included and excluded from the target population and the process, condition, event, or outcome being measured; how to compute the score, etc.
*QDM (formerly called the QDS) elements should be used when available. When quality data elements are needed but are not yet available in the QDM, they will be considered for addition to the QDM.

Table 9: Generic Scale for Rating Subcriterion 2c

Rating	Definition
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

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 decisionmaking. H☐ M☐ L☐ I ☐ Definitions-Table 10

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; H☐ M☐ L☐ I ☐

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. H☐ M☐ L☐ I ☐

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 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. H☐ M☐ L☐ I ☐ Definitions-Table 10

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). H☐ M☐ L☐ I ☐

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. H☐ M☐ L☐ I ☐

4c. Susceptibility to inaccuracies, errors, or unintended consequences and the ability to audit the data items to detect such problems are identified. H☐ M☐ L☐ I ☐

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). H☐ M☐ L☐ I ☐

Note
17. All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.

Table 10: Generic Scale for Rating Usability and Feasibility and Subcriteria

Rating	Definition
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

5. Comparison to Related or Competing Measures Definitions-Table11 Guidance-Figure 1 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. Guidance-Table 13 5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure); Guidance-Table 12 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.

Guidance on Evaluating Related and Competing Measures

For more information, see full report: Guidance for Measure Harmonization.

Table 11: Related versus Competing Measures

	Same concepts for measure focus—target process, condition, event, outcome	**Different concepts for measure focus—target process, condition, event, outcome**
Same target patient population	**Competing measures—Select best measure** from competing measures or justify endorsement of additional measure(s).	**Related measures—Harmonize** on target patient population or justify differences.
Different target patient population	**Related measures—Combine** into one measure with expanded target patient population or justify why different harmonized measures are needed.	Neither harmonization nor competing measure issue

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

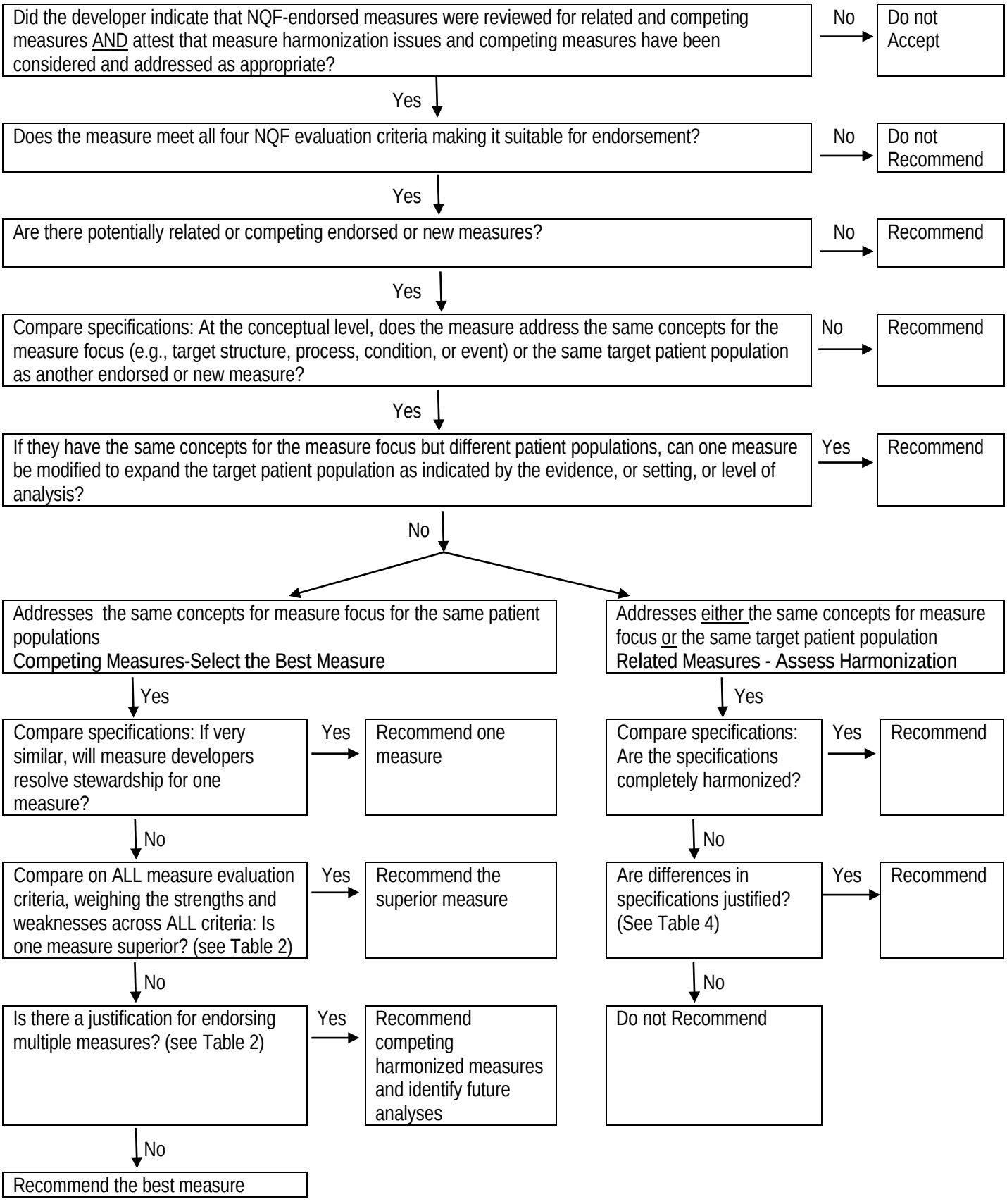

Table 12: Evaluating Competing Measures for Superiority or Justification for Multiple Measures

Steps	Evaluate Competing Measures
1. Determine if need to compare measures for superiority	Work through the steps in the algorithm (Figure 1) to determine if need to evaluate competing measures for superiority (i.e., two or more measures address the same concepts for measure focus for the same patient populations)
2.Assess Competing Measures for Superiority by weighing the strengths and weaknesses across ALL NQF evaluation criteria	Because the competing measures have already been determined to have met NQF's criteria for endorsement, the assessment of competing measures must include weighing the strengths and weaknesses across ALL the criteria and involves more than just comparing ratings. (For example, a decision is not based on just the differences in scientific acceptability of measure properties without weighing the evaluation of importance to measure and report, usability, and feasibility as well.) Impact, Opportunity, and Evidence—Importance to Measure and Report: Competing measures generally will be the same in terms of the measure focus addressing a high-impact aspect of healthcare (1a) and evidence for the focus of measurement (1c). However, due to differences in measure construction, they could differ on alignment with national health goals/priorities or opportunity for improvement. • Compare measures on alignment with national health goals/priorities (1a) • Compare measures on opportunity for improvement (1b) Reliability and Validity—Scientific Acceptability of Measure Properties: • Compare evidence of reliability (2a1-2a2) • Compare evidence of validity, including threats to validity (2b1-2b6) Untested measures cannot be considered superior to tested measures because there would be no empirical evidence on which to compare reliability and validity. (However, a new measure, when tested, could ultimately demonstrate superiority over an endorsed measure and the NQF endorsement maintenance cycles allow for regular submission of new measures.) Compare and identify differences in specifications All else being equal on the criteria and subcriteria, the preference is for: • Measures specified for the broadest application (target patient population as indicated by the evidence, settings, level of analysis) • Measures that address disparities in care when appropriate Usability: • Compare evidence of use and usefulness for public reporting, including availability of data for reporting performance results • Compare evidence of use and usefulness for quality improvement All else being equal on the criteria and subcriteria, the preference is for: • Measures that are publicly reported • Measures with the widest use (e.g., settings, numbers of entities reporting performance results) • Measures that are in use over those without evidence of use Feasibility: • Compare the ease of data collection/availability of required data • Compare the potential for inaccuracies, errors, and unintended consequences All else being equal on the criteria and subcriteria, the preference is for: • Measures based on data from electronic sources

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

Table 13: Sample Considerations to Justify Lack of Measure Harmonization

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