Care Coordination Steering Committee In-Person Meeting

February 28-29, 2012

NATIONAL QUALITY FORUM

1

Welcome and Introductions

Project Introduction

Project Overview

- Phase 1
- Phase 2
  - Performance measure evaluation
    - 15 measures up for endorsement maintenance
  - Review 25 NQF-endorsed Care Coordination Preferred Practices
    - Help promote organizational progress toward better care coordination
    - Help shape measure development goals

Overview of Evaluation Process

Four Major Endorsement Criteria Hierarchy and Rationale

- Describe desirable characteristics of quality performance measures for endorsement
  - **Importance to measure and report:** Measure those aspects with greatest potential of driving improvements; if not important, the other criteria less meaningful (***must-pass***)
  - **Scientific acceptability of measure properties:** Goal is to make valid conclusions about quality; if not reliable and valid, risk of improper interpretation (***must-pass***)
  - **Usable**: Goal is to use for decisions related to accountability and improvement; if not useful, probably do not care if feasible
  - **Feasible: I**deally, cause as little burden as possible; if not feasible, consider alternative approaches
- If suitable for endorsement, evaluate measure harmonization and best-in-class

Evaluation of Already-Endorsed Measures

All measures are expected to meet current criteria and guidance

- Subcriterion 1b (Opportunity for Improvement): Expect data from implementation of the measure
  - Potential for reserve status
- Expanded reliability and validity testing (unless already meet high rating)
- Usability: Actual use in public reporting/other accountability and improvement OR specific plans and timeline
- Feasibility: Problems with implementation or unintended consequences

Generic Rating Scale
1a-High impact, 1b-performance gap, 3-Usability, 4-Feasibility

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)

Low Rating vs. Rating of Insufficient Evidence

- A low rating generally means the evidence/information demonstrates that a criterion is not met
  - For evidence: Depends on combination of quantity, quality, consistency
- Insufficient evidence means either:
  - The evidence does exist and was presented but is not adequate for a definitive answer **OR**
  - The submission was incomplete or deficient in presenting evidence/information that does exist
- Ratings of Low or Insufficient Evidence for a subcriterion result in not passing a criterion but signify different reasons
  - For evidence: Depends on combination of quantity, quality, consistency

1. Importance to Measure and Report

Must-pass criterion: Must meet all 3 subcriteria

1a. High impact

- National health goal or priority
- Data on numbers of persons affected, high resource use, severity of illness, consequences of poor quality

1b. Performance gap/Opportunity for improvement

- Data demonstrating considerable variation in performance OR overall less than optimal performance
- Data on disparities in care
- Potential for reserve status for endorsed measures

1c. Evidence

- Quantity, quality, consistency of body of evidence

Criteria for Reserve Status

Potential Reserve Status for endorsed measures with ***demonstrated high levels of performance***

- The purpose is to retain endorsement of reliable and valid quality performance measures that have overall high levels of performance with little variability so that performance could be monitored in the future if necessary to ensure that performance does not decline
- Exceptional circumstance, not the rule
  - Applies only to highly credible, reliable, and valid measures that have high levels of performance due to quality improvement actions (often facilitated or motivated through public reporting and other accountability programs)
- Additional criteria must be met, so will need to continue evaluation beyond 1b if think might quality

Criteria for Reserve Status

- Evidence for measure focus (1c): Strong direct evidence of a link to a desired health outcome
- For process and structure measures, the measure focus should be proximal to the desired outcome
  - Generally, measures more distal to the desired outcome would not be eligible for reserve status
- Reliability (2a) – high rating
- Validity (2b) – high rating
- The reason for high levels of performance is better performance, not an issue with measure construction
- Demonstrated usefulness for improving quality
- Demonstrated use of the measure

Subcriterion 1c: Submitted vs. Existing Evidence

- Individual committee member preliminary evaluation
  - Rate the measures based on evidence submitted
  - Note if aware of additional evidence
  - Continue to evaluate all remaining criteria
- After workgroup discussion
  - If confident in the evidence presented by committee members **AND** the measure is likely to meet criteria for:
    - High impact (1a), Performance gap (1b) **and**
    - Scientific acceptability of measure properties
      - *Reliability (2a) & Validity (2b)*
  - Could ask developer to provide the additional evidence for consideration

Evidence Rating Scale: Quantity of Body of Evidence

Rating	**Quantity of Body of Evidence:** Total number of studies (not articles or papers)
High	5+ studies
Moderate	2-4 studies
Low	1 study
Insufficient to evaluate	• No empirical evidence OR • Only selected studies from a larger body of evidence

Evidence Rating Scale: Quality of Body of Evidence

Rating	**Quality of Body of Evidence:** Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence
High	RCTs; direct evidence for specific measure focus; adequate size to obtain precise estimates of effect; without serious flaws that introduce bias
Moderate	Non-RCTs w/control for confounders; large, precise estimates of effect **OR** RCTs without serious flaws, but either indirect evidence or imprecise estimate of effect
Low	RCTs w/flaws introduce bias **OR** Non-RCTs w/small or imprecise estimate of effect or without control of confounders
Insufficient to evaluate	• No empirical evidence **OR** • Only selected studies from a larger body of evidence

Evidence Rating Scale: Consistency of Results of Body of Evidence

Rating	**Consistency of Results of Body of Evidence**: 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	Estimates of clinically/practically meaningful benefits & harms to patients consistent in direction & similar in magnitude across preponderance of studies
Moderate	Estimates of benefits & harms consistent in direction but may differ in magnitude (If 1 study then estimate of benefits greatly outweigh harms)
Low	Estimates of benefits & harms differ in both direction and magnitude **OR** wide confidence intervals prevent estimating net benefit (If 1 study then estimate of benefits do not greatly outweigh harms)
Insufficient to evaluate	No assessment of magnitude and direction of benefits and harms to patients

Subcriterion 1c: Evidence Decision Logic

Quantity	Quality	Consistency	Does the measure meet subcriterion 1c?
Moderate or High	Moderate or High	Moderate or High	YES
Low	Moderate or High	Moderate	YES, IF additional research unlikely to change conclusion that benefits to patients outweigh harms. Otherwise NO.
Moderate or High	Low	Moderate or High	YES, IF potential benefits to patients clearly outweigh potential harms. Otherwise NO.
Low, Moderate, or High	Low, Moderate, or High	Low	NO

NOTE: Insufficient evidence – does not pass 1c

Exceptions to the Evidence Subcriterion (1c)

Quantity of Body of Evidence	Quality of Body of Evidence	Consistency of Results of Body of Evidence	Pass Subcriterion 1c
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. Otherwise NO.
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.

2. Scientific Acceptability of Measure Properties

Must-pass criterion: Must meet both subcriteria

2a. Reliability
- 2a1. Precise specifications
- 2a2. Reliability testing—data elements or measure score

2b. Validity (and threats to validity)
- 2b1. Specifications consistent with evidence
- 2b2. Validity testing—data elements or measure score
- 2b3. Justification of exclusions (also relates to evidence)
- 2b4. Risk adjustment
- 2b5. Identification of differences in performance
- 2b6. Comparability of data sources/methods

2c. Disparities – now addressed only in 1b

Reliability and Validity Rating Scales

Rating	Reliability	Validity
High	• Precise specifications; AND • Empirical evidence of reliability of **BOTH** data elements AND measure score	• Specifications consistent w/ evidence; AND • Empirical evidence of validity of **BOTH** data elements AND measure score; AND • Threats to validity empirically assessed and addressed
Moderate	• Precise specifications; AND • Empirical evidence of reliability of **EITHER** data elements OR measure score	• Specifications consistent w/ evidence; AND • Empirical evidence of validity of **EITHER** data elements OR measure score OR systematic assessment of face validity; AND • Threats to validity empirically assessed and addressed

Reliability and Validity Rating Scales

Rating	Reliability	Validity
Low	• Ambiguous specifications; OR • Empirical evidence of unreliability	• Specifications not consistent w/ evidence; OR • Empirical evidence of invalidity; OR • Threats empirically assessed and bias results
Insufficient Evidence	Inappropriate method/scope	• Inappropriate method/scope; OR • Threats not assessed

Evaluation of scientific acceptability of measure properties

Validity Rating	Reliability Rating	Pass *Scientific Acceptability of Measure Properties* for Initial Endorsement*	
High	**Moderate or High**	**Yes**	Evidence of reliability and validity
	Low	No	Represents inconsistent evidence—reliability is usually considered necessary for validity
Moderate	**Moderate or 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.

3. Usability*

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

- 3a. Meaningful, understandable, and useful for public reporting
  - Is it in use for public reporting or an accountability application and if not, what is plan/progress?
  - Is the rationale for use in accountability credible?
- 3b. Meaningful, understandable, and useful for quality improvement
  - Is it in use for improvement, and if not what is the plan/progress?
  - Is the rationale for use in QI credible?

** Currently being revised*

4. Feasibility

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

- 4a. Clinical data generated and used during care process
  - Blood pressure, lab value vs. survey or observation
- 4b. Electronic sources
  - EHR, claims vs. abstracted and entered into database/registry
  - Is there a credible, near-term path to electronic collection?
- 4c. Susceptibility to inaccuracies/ unintended consequences identified
  - Ability to audit and detect?
- 4d. Data collection strategy can be implemented
  - Is it already in operational use or testing indicated ready for operational use?

Criteria for Evaluation – Composite measures

- Individual measures included in a composite must be
  - NQF endorsed; OR
  - Assessed to have met the individual measure evaluation criteria as a first step in evaluating the composite measure

Importance to Measure and Report

- If the component measures meet the criteria 1a, 1b, and 1c, then the composite meets the criteria.
  - A component measure may not be important as an individual measure, but could be an important component of a composite.
- The construct for quality of the composite is clearly described.
  - The component measures are consistent with and representative of the conceptual construct of quality.

Scientific Acceptability of the Measure Properties

- Composite specifications include methods for standardizing scales across component scores, scoring rules, weighting rules, handling of missing data and sample size.
- Reliability testing, validity testing, meaningful differences sub-criteria
- Component analysis demonstrates that the included components fit the conceptual construct.
- Component analysis demonstrates that the included components contribute to the overall variation in the score
- Scoring and weighting rules are consistent with conceptual construct.
- Analysis of missing component effects

Usability and Feasibility

- Usability
  - Data detail is maintained such that the composite can be deconstructed into its components to facilitate transparency and understanding
  - Demonstration that the composite measure achieves the stated purpose (pilot testing or operational data)
- Feasibility
  - Same sub-criteria as for individual measures

5. Comparison to Related or Competing Measures

If a measure meets the four criteria and there are endorsed/new related measures (same measure focus or same target population) or competing measures (both the same measure focus and same target population), the measures are compared to address harmonization and/or selection of the best measure.

- 5a. The measure specifications are harmonized 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.

Competing and Related Measures

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

Addressing Related and Competing Measures

Does the measure meet all four NQF evaluation criteria making it suitable for endorsement?	No →	Do not Recommend
Yes ↓		
Are there potentially related or competing endorsed or new measures?	No →	Recommend
Yes ↓		
Compare specifications: At the conceptual level, does the measure address the same concepts for the measure focus (e.g., target structure, process, condition, or event) or the same target patient population as another endorsed or new measure?	No →	Recommend
Yes ↓		
If they have the same concepts for the measure focus but different patient populations, **can one measure be modified to expand the target patient population as indicated by the evidence, or setting, or level of analysis**?	Yes →	Recommend
No ↓		

Addressing Competing Measures

Addresses the same concepts for measure focus for the same patient populations **Competing Measures → Select the Best Measure**		
↓ Yes		
Compare specifications: If very similar, will measure developers resolve stewardship for one measure?	Yes →	Recommend one measure
↓ No		
Compare on ALL measure evaluation criteria, weighing the strengths and weaknesses across ALL criteria: Is one measure superior? (see Table 2)	Yes →	Recommend the superior measure
↓ No		
Is there a justification for endorsing multiple measures? (see Table 2)	Yes →	Recommend competing harmonized measures and identify future analyses
↓ No		
Recommend the best measure		

Addressing Related Measures for Harmonization

Addresses <u>either</u> the same concepts for measure focus <u>or</u> the same target patient population **Related Measures → Assess Harmonization**		
↓ Yes		
Compare specifications: Are the specifications completely harmonized?	Yes →	Recommend one measure
↓ No		
Are differences in specifications justified? (See Table 4)	Yes →	Recommend the superior measure
↓ No		
Do not Recommend		

Assess for Superiority

- Impact, Opportunity, Evidence—Importance to measure and report
- Reliability and Validity—Scientific Acceptability of Measure Properties
  - Untested measures cannot be considered superior
  - Preference for measures with broadest application and those that address disparities in care
- Usability
  - Preference for measures publicly reported, widest use, in use
- Feasibility
  - Preference for measures based on electronic sources, clinical data from EHRs, freely available

Assess Justification for Multiple Measures

- Value
  - To change to EHR-based measurement
  - Broader applicability if one measure cannot accommodate all patient populations, settings, etc.
  - Increased availability of performance results
- Burden
  - Interpretability across measures
  - Increased data collection
- Does value outweigh burden?

Assess Justification for Lack of Harmonization

- Evidence should guide specifications
- Different data sources may require some differences in technical specifications
- Should not be simply due to proprietary interests or preferences
- The difference does not affect interpretability or burden of data collection
- If it does affect burden, it adds value that outweighs any concern regarding interpretability or burden of data collection

Electronic Voting

Electronic Voting

Committee will vote via hand-held device

- Keypad assigned to each Committee member
  - Automatically on
  - 60-second timer to cast vote
  - Press number on keypad to cast vote
  - Results will appear on the screen
- Voting Response Options:

1 = Yes **2 = No**	**1 = High** **2 = Moderate** **3 = Low** **4 = Insufficient**

Voting Exercise

Did you have any difficulties traveling to Washington, DC?

1=Yes

2=No

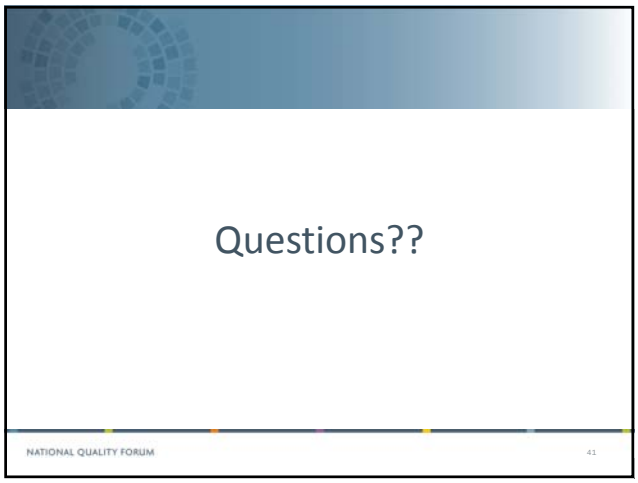
How much snow covers the ground where you live ?

1=Completely

2=Partially

3=Minimally

4=None at all


Questions??
NATIONAL QUALITY FORUM
41