# Measure Testing and Scientific Acceptability of Measure Properties

Evaluation and Measure Submission Guidance

07/12/11



## Measure Evaluation Guidance

- Reports on guidance for measure evaluation:
  - Evidence for the Focus of Measurement and Importance to Measure and Report
  - Measure Testing and Scientific Acceptability of Measure Properties
  - Measure Harmonization
- Updated Measure Evaluation Criteria
- Revised Measure Submission Form
  - Most changes related to guidance on evidence (1c)
  - Some changes related to taxonomy (primarily response options, e.g., setting)
  - Some clarification in wording/instructions

## 2. Reliability and Validity –Scientific Acceptability of Measure Properties

NQF National Quality Forum

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

**2a. Reliability**

- 2a1. Precise specifications (previously 2a) including exclusions (previously 2d)
- 2a2. Reliability testing (previously 2b)—data elements or measure score

**2b. Validity**

- 2b1. Specifications consistent with evidence (new)
- 2b2. Validity testing (previously 2c)—data elements or measure score
- 2b3. Justification of exclusions (previously 2d)—relates to evidence
- 2b4. Risk adjustment (previously 2e)
- 2b5. Identification of differences in performance (previously 2f)
- 2b6. Comparability of data sources/methods (previously 2g)

**2c. Disparities** (previously 2h)



## Measure Testing Guidance: Key Points

NQF National Quality Forum

- Empirical evidence of reliability and validity is expected (measure testing)
- Reliability and validity are demonstrated for the measure as specified (not the measure concept)
- Flexible testing options rather than prescriptive
- Specific thresholds not set – results should be within acceptable norms
- Insufficient evidence cannot be evaluated or considered for endorsement (untested)
- Does not replace need for expertise and judgment
- Strategies to mitigate the burden of testing

## Minimize the Burden of Testing

- Testing of data elements or computed measure score
- Sample
- If empirical evidence of data element validity, separate reliability of data elements not required
- Prior evidence may be used as appropriate
- Face validity accepted (if systematically assessed)



## Notes Reliability & Validity

- Reliability of measure scores primarily assesses amount of variation in scores due to error (noise) vs. true variation (signal)
- Comparing agreement of computed scores from two raters (abstractors) is considered testing at the data element level because the data elements are needed to compute the score
- At the data element level, comparing agreement of data used in a measure with an authoritative source is similar to methods for inter-rater reliability comparing data used in measure from two abstractors; therefore, if data element validity testing conducted, then reliability testing of data elements not required
- Face validity should focus on the measure score

## Measure Testing Resources

- References
- Appendix A  Common Approaches To Measure Testing
- Appendix C  Glossary
  - **Measure testing:** Empirical analysis to demonstrate the reliability and validity of the *measure as specified* including analysis of issues that pose threats to the validity of conclusions about quality of care such as exclusions, risk adjustment/stratification for outcome and resource use measures, methods to identify differences in performance, and comparability of data sources/methods.



## Reliability & Validity Rating Scale

- See Measure Testing Report – Table 2, p.14

| Rating | Reliability | Validity |
|---|---|---|
| **High** | | |
| **Moderate** | | |
| **Low** | | |
| **Insufficient Evidence** | | |

## Evaluation of Scientific Acceptability of Measure Properties

NQF National Quality Forum

| 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.**



## Additional Guidance

NQF National Quality Forum

- Measures specified for EHRs – Measure Testing Report Table 4, p.20
  - Follow same general framework
  - Specific examples for EHRs
- Untested Measures – Measure Testing Report Table 5, p.22
  - Insufficient testing data provided
  - Must be specified to even be eligible
- Endorsement maintenance – Measure Testing Report Table 6, p.23
  - Same criteria
  - Should continue testing until achieves highest ratings

NQF
National Quality Forum

# Measure Submission: Sections 2a2 – 2c Scientific Acceptability of Measure Properties



## Disclaimer

NQF
National Quality Forum

- The following are illustrations of the type of information NQF is seeking on the submission form
  - Not intended as an example for one measure
  - Not intended to represent the only or best approach to measure development and testing
  - Undesirable examples are indicated with an X
- The key points are
  - Provide the information requested
  - Provide substantive information and data in the measure submission form
  - Provide information that demonstrates the criteria are met

## 2a2. Reliability Testing
## 2a2.1. Data/ Sample

NQF National Quality Forum

*(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

Example - Testing at level of data elements

Five group practices

The Medicare claims database

(Note: All the requested information not provided)

Note: The following provides requested information

20 home health agencies representing various types, locations, and sizes

- 4 private, for-profit; 2 public for-profit chain; 6 private nonprofit; 1 health dept.; 5 hospital-based; 2 visiting nurse associations
- Located in 4 states: AZ, MO, NY, TX
- 3 – less than10,000 visits/year; 10 – 10,000-30,000; 7 – greater than 30,000

20-40 patients per agency for a total of 500 patients

Patient case-mix characteristics were similar to national – no significant differences (see report attached, Table 5, p. 20)

Data collected March-June 2009



## 2a2. Reliability Testing
## 2a2.2. Analytic Methods

NQF National Quality Forum

*(Describe method of reliability testing and rationale)*

See attached methodology report

(Note: requested information not provided in form)

Note: The following provides requested information

Inter-rater reliability was assessed for the critical data elements used in this measure because testing conducted for a sample of agencies

Patients were randomly selected from planned visits for start or resumption of care and discharge assessments for each day

2nd nurse assessment within 24 hours

Data analysis included:

- Percent agreement
- Kappa statistic to adjust for chance agreement for categorical data
- ICC for quantitative data

## 2a2. Reliability Testing
## 2a2.3. Testing Results

*(Provide reliability statistics and assessment of adequacy in the context of norms for the test conducted)*

Our expert panel found the measure to be reliable

(Note: Does not provide requested information)

Note: The following provides requested information

**Data Element (N, %Agreement, Kappa)**

Functional status score for ambulation (500, 85%, 0.62)

Functional status score for ambulation prior to this start/resumption of care (495, 83%, 0.55)

Primary diagnosis major diagnostic category (500, 90%, 0.70)

Pain scale (500, 88%, 0.69)

Location prior to this start/resumption of care (500, 91%, 0.72)



## 2b. Validity

2b1.1. Describe how the measure specifications (measure focus, target population, and exclusions) are consistent with the evidence cited in support of the measure focus (criterion 1c) and identify any differences from the evidence.

The evidence demonstrated the association between Hba1c and morbidity and mortality in patients with diabetes. The clinical practice guideline recommended monitoring Hba1c every 3-6 mo (based on expert consensus). The specified measure numerator is counting the number of patients with diabetes who have an annual Hba1c. (Note: Provides the requested information; however in terms of evaluation, the focus is not consistent with evidence presented)

Note: The directness of evidence to the measure as specified provides a foundation for validity of the measure as an indicator of quality

## 2b2. Validity Testing 2b2.1. Data/Sample

*(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

**Example 1** – Face validity at level of measure score

Our expert panel included 20 members including endocrinologists, primary care physicians, nurses, diabetes educators, and patients.

List Members including Name, Credentials, Title, Organization, City, State



## 2b2. Validity Testing 2b2.2. Analytic Method

*(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*

**Example 1** – Face validity at level of measure score

Our expert panel (membership) voted to approve the measure

(Note: Does not provide sufficient information on the method)

Note: The following provides requested information and is focused on measure as specified not just the general idea

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel (membership) was asked to rate their agreement with the following statement:

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

Scale 1-5: 1=Disagree; 3=Moderate Agreement; 5=Agree

## 2b2. Validity Testing
## 2b2.3. Testing Results

NQF National Quality Forum

*(Provide statistical results and assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*

**Example 1** – Face validity at level of measure score

Our expert panel found the measure to be valid

(Note: Does not provide sufficient information on the method)

Note: The following provides requested information

The results of the expert panel rating of face validity (agreement with the statement 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)

N= 20; Mean rating 4.75

Frequency Distribution of Ratings

1 – 0 (Disagree)

2 – 0

3 – 1 (Moderate Agreement)

4 – 3

5 – 16 (Agree)



## 2b2. Validity Testing 2b2.1. Data/Sample

NQF National Quality Forum

*(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

**Example 2** – Validity testing at level of measure score

- 709 hospitals participating in the National Registry of Myocardial Infarction (NRMI) represented various regions, urban/rural location, ownership, teaching/nonteaching, and size
- Included hospitals had at least 12 AMI patients and at least 10 eligible patients for the process measure for a total of xx patients
- 2002-2003 data

## 2b2. Validity Testing
## 2b2.2. Analytic Method

NQF National Quality Forum

*(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*

**Example 2** – Validity testing at level of measure score

See attached methodology report

(Note: Requested information not provided in form)

Note: The following provides requested information

- Validity testing of the hospital score on the process measure of timely reperfusion in AMI patients was conducted by correlation analysis to the outcome of 30-day mortality.
- The risk standardized 30-day mortality rate using a hierarchical generalized linear model (HGLM)
- Various secondary and sensitivity analyses to help interpret results
- See attached published report: Bradley EH, Herrin J, Elbel B, et al., Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality, *JAMA,* 2006;296(1):72-78.



## 2b2. Validity Testing
## 2b2.3. Testing Results

NQF National Quality Forum

*(Provide statistical results and assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*

**Example 2** – Validity testing at level of measure score

- Timely reperfusion therapy N=709 hospitals; mean=54.5 (SD=13.3); 25$^{th}$ percentile=45.5; median=53.9; 75$^{th}$ percentile=63.9
- Correlation coefficient between hospital rates for timely reperfusion and 30-day mortality = -0.18 (p<.001)
- Although correlation is significant and in the hypothesized direction, it is small and timely reperfusion accounts for only 3.3% of the variability in risk standardized 30-day mortality
- To facilitate interpretation, analyses demonstrated that a composite of 5 AMI medication process measures accounted for 6% of variation, teaching status explained 6.5% of variation, case volume 6.8%, geographical variation 4.5%
- Although this one process measure score for timely reperfusion cannot be used alone to infer mortality, the results do not negate the importance of continuing to measure given the strong evidence base and until research identifies process performance measures with stronger links to outcomes

## 2b3. Exclusions 2b3.1. Data/sample for Analysis of Exclusions

NQF National Quality Forum

*(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

- 50 hospitals representing various types, locations, and sizes
  - 5 public; 20 community; 10 teaching; 5 rural; 10 for-profit
  - Located in 10 states: AZ, CA, FL, MI, MO, NY, NV, OH, TX
  - 20 >500 beds; 20 250-500 beds; 10 < 250 beds
- 20-40 patients per hospital for a total of 1500 patients
- Patient case-mix characteristics were similar to national – no significant differences (see report attached, Table 5, p. 20)
- Data collected January-May 2010



## 2b3. Exclusions 2b3.2. Analytic Method

NQF National Quality Forum

*(Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*

- One exclusion was not specifically indicated in the evidence for influenza immunization
- The exclusion for leaving the hospital against medical advice was analyzed for frequency and variability across providers

## 2b3. Exclusions 2b3.3. Results

NQF National Quality Forum

*(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses)*

Hospitalizations in which the patient left AMA accounted for 1.2% of the hospitalizations

The percentile distribution was:

| | |
|---|---|
| 10th percentile | 0.9% |
| 25th percentile | 1.0% |
| 50th percentile | 1.2% |
| 75th percentile | 1.5% |
| 90th percentile | 2.0% |



## 2b4. Risk Adjustment 2b4.1. Data/sample

NQF National Quality Forum

*(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

Note: Risk model validation is in addition to validity testing addressed in 2b2 (data element/score).

Medicare administrative datasets that contain HF FFS hospitalizations for patients discharged in 2003 and 2004. The datasets also contain administrative data for each patient in the year before each index admission and the 30 days following the index admission. All hospitals are included.

Medicare Part A inpatient claims

Hospital outpatient data – 12 months pre-index admission

Part B data – 12 months pre-index admission

## 2b4. Risk Adjustment
## 2b4.2. Analytic Method

NQF National Quality Forum

*(Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*

The risk model is derived using a randomly selected half of the hospitalizations in 2004 ("derivation sample"). The performance of the model is then evaluated using patients contained in the other half of the dataset. We compute indices that describe model performance in terms of predictive ability, discriminant ability, and overall fit. We assess variability over time using 2003 data.

We derive the model using risk factor variables that exclude potential complications. To consolidate the 15,000+ ICD-9-CM codes into clinically coherent groupings, we use the Condition Categories (CCs) from CMS's Hierarchical Condition Category (HCC) methodology, a publicly available diagnostic grouping system (Pope et al., 2000). The final risk adjustment variables were selected by a team of physicians and analysts primarily based on their clinical relevance but with knowledge of their strength of association with the readmission outcome using 200 bootstrap samples.



## 2b4. Risk Adjustment
## 2b4.3. Testing Results

NQF National Quality Forum

*(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*

Of 99 initial candidate variables, 37 were retained in the final model. 25 were associated with readmission $p < 0.001$ in 70% of bootstrap samples. The others were included if 1)considered markers for frailty/end of life, 2) might have disproportionate share of patients (e.g., cancer), or 3) on the same clinical spectrum as a variable above the 70% cutoff and were clinically important for HF patients (e.g. asthma and COPD and depression and other psychiatric disorders)

## 2b4. Risk Adjustment
## 2b4.3. Testing Results cont.

NQF National Quality Forum

The derivation model has modest discrimination ($R^2$ = 0.034), calibration, and fit. The patient-level predicted readmission rate ranges from 15% in the lowest predicted decile to 37% in the highest predicted decile, a range of 22%. The area under the ROC curve is 0.601. For comparison, a model with age and gender had an ROC of 0.516 and a model with all candidate variables had an ROC equal to 0.604.

The standardized regression coefficients and standard errors for the 2004 validation dataset are shown in Table 9, and the performance metrics are shown in Table 11. The performance was not substantively different in this validation sample ($R^2$ = 0.04 and ROC area = 0.60).

See methodology report attachment, p.x for risk decile plots; p.x for Table 9 and p.x for Table 11.



## 2b4. Risk Adjustment
## 2b4.3. Testing Results cont.

NQF National Quality Forum

The discrimination and the explained variation of the model at the patient-level are consistent with the few published models of readmission after HF that report predictive ability (Philbin and DiSalvo, 1999; Yamokoski et al, 2007). We excluded covariates such as potential complications, certain patient demographics (e.g., race, socioeconomic status), and patients' admission path and discharge disposition (e.g. admit from, or discharge to, a skilled nursing facility). These characteristics may be associated with readmission and thus could increase the model performance to predict patient readmissions. However, these variables may be related to quality or supply factors that should not be included in an adjustment that seeks to control for patient clinical characteristics.

## 2b4. Risk Adjustment

NQF National Quality Forum

2b4.4. If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment

Note: Any outcome measure (intermediate clinical outcome or health outcome should have an assessment of whether risk adjustment is needed for fair comparisons across providers)

Are there potential patient characteristics (at start of care) that influence achievement of the outcome?
Are they statistically significantly associated with the outcome of interest?
Does the distribution of patients with those characteristics vary across providers?



## 2b5. Differences 2b5.1. Data Sample

NQF National Quality Forum

*(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

Provide specific information on the data or sample as in the reliability testing example (2a2.1)

## 2b5. Differences 2b5.2. Analytic Method

NQF National Quality Forum

*(Describe methods and rationale to identify statistically significant and practical/meaningful differences in performance)*

A confidence interval was computed for each provider's score and if it did not contain the average, the provider is identified as better or worse than average



## 2b5. Differences 2b5.3. Results

NQF National Quality Forum

*(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*

Not applicable

Note: If any testing has been done, then performance scores should be computed and reported for the entities included in testing.

Scores on this measure: N=1000, Mean 95%, SD 9.0
10th percentile – 87%
25th percentile – 94%
50th percentile – 98%
75th percentile – 100%
90th percentile - 100%

Of the 1000 providers, 1% were statistically significantly better than average and 5% worse than average

## 2b6. Comparability of Multiple Data Sources/Methods

NQF National Quality Forum

**2b6.1. Data Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*

Same detail as above

**2b6.2. Analytic Method** *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*

e.g., Correlation analysis, analysis of rank orders

**2b6.3. Testing Results** *(Provide statistical results (e.g., correlation statistics, comparison of rankings) and assessment of adequacy in the context of norms for the test conducted)* Provide substantive results



## 2c. Disparities in Care

NQF National Quality Forum

**2c.1. If measure is stratified for disparities, provide stratified results** *(Scores by stratified categories/cohorts)*

Note: This is for scores on the specific measure under consideration (not from studies or other data, which should be reported under 1b)

**2c.2. If disparities have been reported/identified, but measure is not specified to detect disparities, please explain.**

Not applicable

Note: If no disparities have been identified, that should be stated

## Generic Rating Scale

NQF
National Quality Forum

- **Used with 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) |