

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b6)

Measure Title: Heart Failure (HF): Left Ventricular Function (LVF) Testing

Date of Submission: [12/23/2013](#)

Type of Measure:

<input type="checkbox"/> Composite – STOP – use composite testing form	<input type="checkbox"/> Outcome (including PRO-PM)
<input type="checkbox"/> Cost/resource	<input checked="" type="checkbox"/> Process
<input type="checkbox"/> Efficiency	<input type="checkbox"/> Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.
- For outcome and resource use measures, section 2b4 also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), section 2b6 also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*including questions/instructions*; minimum font size 11 pt; do not change margins). **Contact NQF staff if more pages are needed.**
- Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](#).

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing ¹⁰ 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.

2b2. Validity testing ¹¹ 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; ¹²

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). ¹³

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 patient 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; ^{14,15} 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 ¹⁶ 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.

Notes

10. 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).

11. 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.

12. 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.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. 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.

16. 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.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing, (e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? *(Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. **If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.**)*

Measure Specified to Use Data From: (must be consistent with data sources entered in S.23)	Measure Tested with Data From:
<input type="checkbox"/> abstracted from paper record	<input type="checkbox"/> abstracted from paper record
<input type="checkbox"/> administrative claims	<input type="checkbox"/> administrative claims
<input checked="" type="checkbox"/> clinical database/registry	<input checked="" type="checkbox"/> clinical database/registry
<input type="checkbox"/> abstracted from electronic health record	<input type="checkbox"/> abstracted from electronic health record
<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs	<input type="checkbox"/> eMeasure (HQMF) implemented in EHRs
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset *(the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).*

Clinical registry data

1.3. What are the dates of the data used in testing? 1/1/2011 – 12/31/2012

1.4. What levels of analysis were tested? *(testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan)*

Measure Specified to Measure Performance of: (must be consistent with levels entered in item S.26)	Measure Tested at Level of:
<input checked="" type="checkbox"/> individual clinician	<input checked="" type="checkbox"/> individual clinician
<input checked="" type="checkbox"/> group/practice	<input checked="" type="checkbox"/> group/practice
<input type="checkbox"/> hospital/facility/agency	<input type="checkbox"/> hospital/facility/agency
<input type="checkbox"/> health plan	<input type="checkbox"/> health plan
<input type="checkbox"/> other: Click here to describe	<input type="checkbox"/> other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? *(identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)*

Registry data consisting of providers with eligible patient populations in 2011 and 2012 were obtained. In 2011 199 providers had eligible patients with 63 submitting the measure for one or more patients. In 2012 174 providers had eligible patients with 40 submitting data for the measure.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? *(identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample)*

199 providers had 2928 eligible patients in 2011. Data were 963 patients were submitted to the registry by 63 providers.

In 2012 174 providers had 2119 eligible patients and 40 submitted data for 367 patients to the registry. This was the total population reporting this measure.

Provider and patients demographics were unavailable for this analysis.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

For testing we selected providers with at least 10 cases for each reporting year (not necessarily in both years). In 2011 there were 41 providers reporting 798 cases and in 2012 there were 13 providers reporting 283 cases.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? *(may be one or both levels)*

☐ **Critical data elements used in the measure** *(e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements)*

☒ **Performance measure score** *(e.g., signal-to-noise analysis)*

2a2.2. For each level checked above, describe the method of reliability testing and what it tests *(describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used)*

Reliability was calculated according to the methods outlined in a technical report prepared by J.L. Adams titled “The Reliability of Provider Profiling: A Tutorial” (RAND Corporation, TR-653-NCQA, 2009). In this context, reliability represents the ability of a measure to confidently distinguish the performance of one physician from another. As discussed in the report: “Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error.”

According to this approach, reliability is estimated with a beta-binomial model. The beta-binomial model is appropriate for measuring the reliability of pass/fail measures such as those proposed.

2a2.3. For each level checked above, what were the statistical results from reliability testing? (e.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis)

Between clinic variance: 0.0036

2011 Registry Data					
Practice	Number of Patients	Performance measure rate	Reliability	95% CI	
7004684061	17	76.5%	0.25	0.22	- 0.28
7014273047	31	90.3%	0.56	0.54	- 0.58
7064649163	13	84.6%	0.26	0.23	- 0.29
7094753382	21	100.0%	1.00	0.97	- 1.00
7105695072	24	91.7%	0.53	0.50	- 0.55
7115245161	25	88.0%	0.46	0.43	- 0.48
7125919629	29	86.2%	0.46	0.44	- 0.49
7185405490	19	94.7%	0.58	0.55	- 0.60
7195846888	12	91.7%	0.36	0.32	- 0.39
7220761804	11	90.9%	0.32	0.29	- 0.36
7260979415	33	93.9%	0.67	0.65	- 0.69
7328947979	11	100.0%	1.00	0.96	- 1.00
7338156545	14	78.6%	0.23	0.20	- 0.26
7358010964	21	100.0%	1.00	0.97	- 1.00
7358014196	11	90.9%	0.32	0.29	- 0.36
7378499542	20	95.0%	0.60	0.57	- 0.63
7388579950	15	100.0%	1.00	0.97	- 1.00
7419176895	15	100.0%	1.00	0.97	- 1.00
7469279115	17	76.5%	0.25	0.22	- 0.28
7507145741	38	100.0%	1.00	0.98	- 1.00
7527987622	13	100.0%	1.00	0.97	- 1.00
7603106147	25	60.0%	0.27	0.25	- 0.29
7623751471	22	86.4%	0.40	0.37	- 0.42
7693851210	29	79.3%	0.39	0.36	- 0.41
7693974289	35	82.9%	0.47	0.45	- 0.49
7716244812	17	82.4%	0.29	0.27	- 0.32
7736097388	19	94.7%	0.58	0.55	- 0.60
7736185781	21	95.2%	0.62	0.60	- 0.65
7776896799	21	85.7%	0.38	0.35	- 0.40
7822762196	10	90.0%	0.28	0.25	- 0.32
7872401659	34	97.1%	0.81	0.79	- 0.83
7872466161	10	100.0%	1.00	0.96	- 1.00
7872466844	13	92.3%	0.39	0.36	- 0.43
7911230172	27	96.3%	0.73	0.71	- 0.75
7911268193	16	81.3%	0.27	0.24	- 0.30
7921721402	13	92.3%	0.39	0.36	- 0.43
7931135212	16	93.8%	0.49	0.46	- 0.52
7951038578	16	87.5%	0.34	0.31	- 0.37
7981537692	30	96.7%	0.77	0.75	- 0.79
7991721608	63	98.4%	0.93	0.92	- 0.95
7991816181	30	96.7%	0.77	0.75	- 0.79
Median		92.30%	0.49	0.47	- 0.50

Between clinic variance: 0.0038

2012 Registry Data				
Practice	Number of Patients	Performance measure rate	Reliability	95% CI
7014632018	14	100.0%	1.00	0.97 - 1.00
7094753382	15	100.0%	1.00	0.97 - 1.00
7328947979	13	100.0%	1.00	0.97 - 1.00
7378854330	10	100.0%	1.00	0.96 - 1.00
7398966189	13	100.0%	1.00	0.97 - 1.00
7409020257	10	80.0%	0.19	0.15 - 0.23
7439304758	25	100.0%	1.00	0.98 - 1.00
7507145741	26	100.0%	1.00	0.98 - 1.00
7527987622	10	90.0%	0.29	0.26 - 0.33
7812148628	31	100.0%	1.00	0.98 - 1.00
7872401659	24	95.8%	0.69	0.67 - 0.72
7901171523	37	100.0%	1.00	0.98 - 1.00
7991721608	59	100.0%	1.00	0.98 - 1.00
Median		100.0%	1.00	0.98 - 1.00

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.e., what do the results mean and what are the norms for the test conducted?)

The 2011 population of practices shows a varying degree of reliability (scores fell between 0.23 to 1.0) with a median .49 reliability score indicating fair ability to determine differences between practices.

The 2012 population of practices who reported this measure in the registry decreased significantly. The median reliability statistic of 1.0 demonstrates perfect reliability.

The small number of patients and the nearly uniform 100% performance rate in the 2012 population makes it difficult to draw strong conclusions. However, given the available data, the practice specific reliability based on between clinic variance shows a fair level of reliability in 2011 and almost perfect reliability in 2012.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (may be one or both levels)

☐ Critical data elements (data element validity must address ALL critical data elements)

☒ Performance measure score

☐ Empirical validity testing

☒ Systematic assessment of face validity of performance measure score as an indicator of quality or resource use (i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance)

2b2.2. For each level checked above, describe the method of validity testing and what it tests (describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)

Quality Insights of Pennsylvania conducts an Environmental Scan to evaluate the most current research and evidence-based guidelines. The TEP, composed of subject matter specialists and experts with technical measure expertise evaluates the results of the review and provides recommendations based on the

scientific merits of the evidence using the Strength of Recommendation Taxonomy (SORT). The TEP also reviews and establishes the measure's ability to capture what it is designed to capture using a consensus process.

The initial measure development process included alpha-testing in the field with select providers and a public comment period. During the Reliability Testing, Quality Insights again convened a TEP for Environmental Scan review as well as a detailed analysis of beta testing results. Based on the process of multiple stakeholder input, expert panel discussion and public comment, face and content validity of CMS/Quality Insights measures can be assumed to be established.

2b2.3. What were the statistical results from validity testing? (e.g., correlation; t-test)

N/A

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.e., what do the results mean and what are the norms for the test conducted?)

Based on the process of multiple stakeholder input, expert panel discussion and public comment, face and content validity of CMS/Quality Insights measures can be assumed to be established.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — skip to section 2b4

2b3.1. Describe the method of testing exclusions and what it tests (describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used)

Numbers of exclusions reported to registry in two performance periods were assessed for overall impact on performance scores.

2b3.2. What were the statistical results from testing exclusions? (include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores)

Of providers with 10 or more cases the rate of exclusions was 4.7% in 2011 and 4.5% in 2012.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (i.e., the value outweighs the burden of increased data collection and analysis. *Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Instances of reported exclusions were relatively small and include patient refusal and urgent or emergent situations where delay of treatment would jeopardize the patient's health status.

—

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section [2b5](#).

2b4.1. What method of controlling for differences in case mix is used?

- ☒ **No risk adjustment or stratification**
- ☐ **Statistical risk model with** [Click here to enter number of factors](#) **risk factors**
- ☐ **Stratification by** [Click here to enter number of categories](#) **risk categories**
- ☐ **Other,** [Click here to enter description](#)

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

n/a

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of $p < 0.10$; correlation of x or higher; patient factors should be present at the start of care and not related to disparities)

n/a

2b4.4. What were the statistical results of the analyses used to select risk factors?

n/a

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (describe the steps—do not just name a method; what statistical analysis was used)

n/a

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

if stratified, skip to [2b4.9](#)

**2b4.6. Statistical Risk Model Discrimination Statistics (e.g., c-statistic, R-squared): **

n/a

2b4.7. Statistical Risk Model Calibration Statistics (e.g., Hosmer-Lemeshow statistic):

n/a

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

n/a

2b4.9. Results of Risk Stratification Analysis:

n/a

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., what do the results mean and what are the norms for the test conducted)

n/a

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b*)

Variation of performance rates was analyzed to determine central tendency, standard deviation and quartile values.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (*e.g., number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2011 Registry Data	
N (practices with >10 cases)	41
Mean performance score	90.70%
Standard deviation	0.086
Max score	100%
90th percentile	100%
75th percentile	96.70%
Median	92.30%
25th percentile	86.20%
10th percentile	79.30%

2012 Registry Data	
N (practices with >10 cases)	13
Mean performance score	97.40%
Standard deviation	0.059
Max score	100%
90th percentile	100%
75th percentile	100%
Median	100%
25th percentile	100%
10th percentile	90.00%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.e., *what do the results mean in terms of statistical and meaningful differences?*)

The small number of practices submitting cases to the registry makes it difficult to determine meaningful differences in performance rate. We cannot draw conclusions about the population of eligible providers as it is unlikely that these practices represent the population of eligible providers.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). **If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (e.g., *correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.e., *what do the results mean and what are the norms for the test conducted*)