

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*): 2450

Measure Title: HF: Symptom and Activity Assessment

Date of Submission: 12/23/2013

Type of Measure:

<input type="checkbox"/> Composite – <i>STOP – use composite testing form</i>	<input type="checkbox"/> Outcome (<i>including PRO-PM</i>)
<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 testing¹⁰ 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. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing¹¹ 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. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

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 decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (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.

2b7. For eMeasures, composites, and PRO-PMs (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

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).

The primary analysis was performed at the physician level, with each heart failure (HF) encounter being considered a measurement opportunity for the physician. We therefore used all patient encounters for patients with Heart Failure (HF) in the PINNACLE Registry that occurred during the one-year study period. Providers with fewer than 10 HF patient encounters during the study period were excluded, since estimates of reliability are unstable with such small numbers. In addition, practices where no New York Heart Association Classification was documented for any HF patient encounter during the study period were also excluded, as it was assumed that there was a systematic coding error in these practices.

1.3. What are the dates of the data used in testing? The primary analysis included encounters between 1/1/2012-12/31/2012. Additionally we used data from 1/1/2011 thru 12/31/2011 for temporal comparisons.

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 type="checkbox"/> group/practice	<input 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)

1270 providers participating in the PINNACLE registry in 2012 met the minimum number of eligible visits (10) in 2012 for inclusion in the primary analysis. The average number of eligible visits for these providers was 425. The range of number of visits for included providers is 5059 to 10. A description of provider characteristics is provided below:

	Total n = 1270
Provider gender	
(1) Male	1003 (79.2%)
(2) Female	264 (20.8%)
Missing (.)	3
Provider categories	
NP/PA	138 (11.0%)
MD/DO	1074 (85.9%)
RN/nurses	38 (3.0%)
Missing (.)	20
Region	
(1) Northeast	207 (16.3%)
(2) Midwest	408 (32.1%)
(3) South	436 (34.3%)
(4) West	219 (17.2%)

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)*

There were 539,430 patient visits in 2012 included in the primary analysis. The table below provides a summary of patient characteristics.

	Total n = 539430
--	-----------------------------

	Total
	n = 539430
Race	
(1) White	328264 (87.2%)
(2) Black	39865 (10.6%)
(3) Other	8444 (2.2%)
Missing (.)	162857
Insurance	
(0) No insurance	31177 (6.3%)
(1) Private	272809 (55.4%)
(2) Medicare	174165 (35.4%)
(3) Medicaid	10854 (2.2%)
(4) Other	3354 (0.7%)
Missing (.)	47071
Age	
18 to <60	108378 (20.1%)
60 to <70	129264 (24.0%)
70 to <80	153026 (28.4%)
80 to 112	148762 (27.6%)
Sex	
(1) Male	297783 (55.2%)
(2) Female	241608 (44.8%)
Missing (.)	39
BMI	30.0 ± 7.0
Missing	128702
Diabetes	192369 (35.7%)
CAD	402086 (74.5%)
Hypertension	489445 (90.7%)
AFib	228161 (42.3%)
HF	539430 (100.0%)
PAD	131347 (24.3%)
Prior Stroke/TIA	147720 (27.4%)
MI history	169501 (31.4%)

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.

The dataset described above was used for all aspects of testing.

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 of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error)], where the latter represents the within-physician estimate of our error in assessing their ‘true’ performance. Using this analytic approach, the reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician’s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

Reliability is estimated across five different points: at the minimum number of quality reporting events for the measure (all providers); above the mean number of quality reporting events per physician; and above the 25th, 50th and 75th percentiles of the number of quality reporting events.

2a2.3. For each level of testing 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)

2011 – In 2011, the signal-noise ratios are shown below:

Description	Number of Patients	Signal-to-Noise Ratio
Minimum	10	0.992
25th percentile	100	0.996

Description	Number of Patients	Signal-to-Noise Ratio
50th percentile	217	0.998
75th percentile	449	0.999
Average	362	0.999

2012 – In 2012, the signal-noise ratios are shown below:

Description	Number of Patients	Signal-to-Noise Ratio
Minimum	10	0.994
25th percentile	141	0.997
50th percentile	280	0.998
75th percentile	508	0.999
Average	425	0.999

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?)

For this measure the reliability was very high and was similar for 2011 and 2012, supporting the reproducibility of the estimates across annual reporting periods. At the minimum number of patient visits required (>10) the average reliability was 0.992 and 0.994, respectively, for 2011 and 2012. For providers above the median number of patient encounters, the reliability was even higher at 0.998 in both years.

A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to actual differences in performance. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability. This measure has excellent reliability regardless of the number of patient encounters included in the analysis. This suggests that for physicians with an average or greater number of events the measure has excellent reliability.

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 of testing 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)*

Content validity for this measure was systematically assessed by expert work group members during the development process during extensive discussion and a final confidential vote. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and concurrent formal peer review process. Additionally, comments were solicited from a panel of consumer, purchaser, and patient representatives convened by the AMA-PCPI specifically for this purpose. All comments received were reviewed by the expert work group and the measures were adjusted as needed. Additionally, the measure underwent review and approval by the Board of Trustees of the ACC and the Science Advisory and Coordinating Committee of the AHA, as well as review and voting by the PCPI membership. Members of the expert work group that developed the measure included: Robert O. Bonow, MD, MACC, FAHA, MACP (Co-Chair) (cardiology); Theodore G. Ganiats, MD (Co-Chair) (family medicine; measure methodology); Craig T. Beam, CRE (patient representative); Kathleen Blake, MD (cardiac electrophysiology); Donald E. Casey, Jr., MD, MPH, MBA, FACP, FAHA (internal medicine); Sarah J. Goodlin, MD (geriatrics, palliative medicine); Kathleen L. Grady, PhD, APN, FAAN, FAHA (cardiac surgery); Randal F. Hundley, MD, FACC (cardiology, health plan representative); Mariell Jessup, MD, FACC, FAHA, FESC (cardiology, heart failure); Thomas E. Lynn, MD (family medicine, measure implementation); Frederick A. Masoudi, MD, MSPH (cardiology); David Nilasena MD, MSPH, MS (general preventive medicine, public health, measure implementation); Ileana L. Piña, MD, FACC (cardiology, heart failure); Paul D. Rockswold, MD, MPH (family medicine); Lawrence B. Sadwin (patient representative); Joanna D. Sikkema, MSN, ANP-BC, FAHA (cardiology); Carrie A. Sincak, PharmD, BCPS (pharmacy); John Spertus, MD, MPH (cardiology); Patrick J. Torcson, MD, FACP, MMM (hospital medicine); Elizabeth Torres, MD (internal medicine); Mark V. Williams, MD, FHM (hospital medicine); John B Wong, MD (internal medicine).

Content Validity for this measure is exceedingly strong given that HF is an incurable condition in which the primary goals of therapy are to improve survival and minimize the symptoms and functional limitations of the disease. There are a myriad of strategies for improving the health status (symptoms, function and quality of life) of HF patients. These include medications (e.g. isosorbide dinitrate and hydralazine), devices (e.g. biventricular pacing, ventricular assist devices) and cardiac transplantation. The relevance of these therapies, and their potential benefit, are often directly associated with the health status of patients. While patient-reported outcomes, such as the Kansas City Cardiomyopathy Questionnaire or Minnesota Living with Heart Failure Questionnaire, might be considered better assessments of patients' health status (and would meet the definition of this measure), they are seldom used in routine clinical care. However, the New York Heart Association is a clinician-reported outcome that has been in use since 1954 and is readily reportable, with minimal effort, by clinicians caring for HF patients. Given the importance of patients' health status (symptom control and function) in clinically managing HF, reporting any symptom and functional assessment is a guideline-endorsed standard of care. Ultimately, the actual health status of patients is an important and relevant outcome that, once risk-adjusted, will lay the foundation for comparing the quality of disease management by providers.

However to build the requisite risk-adjustment models, it is critically important to have large datasets with these outcomes. Thus, for the time being, we believe that the mere assessment and recording of patients' health status (which is deplorably low, as shown later in this document) is an important step forward in advancing the quality agenda for HF patients.

Face validity of the measure score was systematically assessed as follows:

After the measure was fully specified, members of two existing committees, one at the ACC, one at AHA, and one joint ACC/AHA, with expertise in in general cardiology, interventional cardiology, heart failure, electrophysiology and quality improvement, outcomes research, informatics and performance measurement, who were not involved in development of the measure, were asked to review the measure specifications and 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."

The respondents recorded their rating on a scale of 1-5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5= Strongly Agree

There were 17 committee members who completed the survey. One respondent was excluded because he was a member of the work group that developed the measures. Further information on the survey respondents is available if needed.

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

The results of the expert panel rating of the validity statement were as follows:

N = 16; Mean rating = 3.75 and 63% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 3

3 - 3 (Neither Agree nor Disagree)

4 - 5

5 - 5 (Strongly Agree)

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?)

The measure was judged to have moderate to high face validity by both its clinical importance and the group of experts asked to rate it. The majority of experts agreed that the measure, as specified, will provide an accurate reflection of quality and can be used to distinguish good and poor quality.

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)

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)

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*)

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.

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*)

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

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*)

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*):

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

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

2b4.9. Results of Risk Stratification Analysis:

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*)

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 that were assessed*)

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

We examined variation in provider performance on this measure based on sex, age, race and a number of other patient factors to identify variations. Full testing report with information on all patient characteristics tested is available in Appendix A-1. In Summary, there were no clinically significant differences in practitioners' reporting of patients' health status, with the largest 'disparity' being that black patients (40.4%) were slightly more likely than white (37.5%) and other races (32.7%) to have their health status recorded.

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*)

We observed enormous variability in the frequency with which providers reported the symptoms and function of their HF patients, ranging for 0% to 100%. The distributions of health status assessment are shown below:

2011

# of providers	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Std Dev
1262	0.00%	0.43%	36.8%	79.6%	100%	39.3%

2012

# of providers	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Std Dev
1270	0.00%	1.06%	35.3%	73.3%	100%	37.5%

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?)

We observed extraordinary variation in this measure across providers caring for HF patients, ranging from providers who assessed functional status in none of their patients' visits to others who assessed it in every patient visit. This was observed in both 2011 and 2012. We not only describe the distribution of performance, but also summarize this variation by calculating the median rate ratio (MRR). The MRR comes from a hierarchical model that adjusts for patient characteristics and examines the variation in the likelihood that one physician versus another would have assessed the patient's symptoms and function. This can be thought of as the likelihood that a statistically identical patient, presenting to 2 different providers in our sample, would have had their health status assessed.

In 2011: Enormous variability was noted among providers. The performance-met rate range was 0-100% with the inter-quartile range being 0.4% to 79.6%. This yielded a Median Rate Ratio of 9.03(8.21, 10.02). The Median Rate Ratio measures the variation between clusters by comparing 2 persons from two randomly chosen different clusters and an MRR of 9.03 indicates an enormous amount of variation among providers with a statistically identical patient being seen by 2 random providers in PINNACLE having, on average, a 9-fold greater likelihood of having their symptoms and function being documented by one provider as compared with another.

In 2012: A large variability was noted among providers. The performance-met rate range was 0-100% with the inter-quartile range being 1.1% to 73.3%. This yielded a Median Rate Ratio of 8.34 (95% CI=7.62, 9.22). As for 2011, an MRR of 8.34 indicates an enormous amount of variation among the providers, with a statistically identical patient being seen by 2 random providers in PINNACLE having, on average, a 8.3-fold greater likelihood of having their symptoms and function being documented by one provider as compared with another..

Given the clinical importance of clearly documenting patients' symptoms and function, coupled with the wide inter-provider variability and very high reliability of this measure (signal-to-noise ratios >0.99), we believe that this measure is exquisitely capable of detecting providers with better or worse quality of HF care. The minimal differences in performance by patient characteristics makes sense given that this is a provider, rather than a patient, behavior and the variations detected are likely due to real differences in performance between 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 data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

This section is not relevant, given that a single specification set and data source were 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*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

In PINNACLE, if a data field is not fulfilled, it is assumed that the process of care was not done. Thus, there are no missing values for this measure, although it is conceivable that the physician did assess the symptoms and function of the patient, but did not record them in the electronic health record, from which the data in PINNACLE are directly abstracted. If there are ‘missing’ assessments, we believe that these will be rapidly corrected once there is physician-level accountability for recording these data occurs.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Given our assumptions, noted above, we did not conduct an empirical analysis of the frequency or distribution of missing data. For this measure, missing data represents a failure.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders)

and how the specified handling of missing data minimizes bias? (i.e., *what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Our assumption is that there is no missing data.