

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

Measure Number (if previously endorsed): 0079

Measure Title: Heart Failure: Left Ventricular Ejection Fraction Assessment (Outpatient Setting)

Date of Submission: 6/23/2015

Type of Measure:

☐ Composite – STOP – use composite testing form	☐ Outcome (including PRO-PM)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ 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](#).
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

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. 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 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 (including clinical and sociodemographic factors) that influence the measured outcome 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. 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:
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ 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 conducted at the level of the individual provider and included all patients with heart failure (HF) cared for by that provider and captured in the PINNACLE Registry during the one-year study period. The PINNACLE Registry systematically maps each practice's Electronic Health Record to the data elements required for the Registry, with careful validation of the translation process prior to enrollment and reporting the results back to the practice. Using these data, we were able to calculate the number of patients who should have undergone a left ventricular ejection fraction (LVEF) assessment during the 12-month period, or a strong reason why the LVEF assessment was not performed is documented. This means that every patient in that provider's practice is included. For this measure, providers with less than 10 eligible patient encounters during the study period were excluded, since estimates of reliability are unstable with such small numbers. All other cases from all practices and providers were included. We included all visits for each patient in these analyses and meeting the performance measure on any single visit within the year met the criterion for this measure.

1.3. What are the dates of the data used in testing? [Click here to enter date range](#)

The primary analysis included encounters between 01/01/2014-12/31/2014. Additionally, we used data from 01/01/2013 thru 12/31/2013 for temporal comparison.

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:
☒ individual clinician	☒ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency

☐ health plan	☐ health plan
☐ other: Click here to describe	☐ 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)*

2013

2,254 providers met the minimum number of eligible patients (10) for inclusion in the reliability analysis. The average number of eligible patients for providers included is 181.6 for a total of 409,332 patients. The range of number of patients for providers included is from 10 to 2,730. As noted above, providers with fewer than 10 eligible patient encounters during the study period were excluded.

The unit of analysis for this measure is the provider. A description of the providers studied for the 2013 calendar year is shown below:

	Total
	n = 2254
Provider gender	
(1) Male	1819 (80.8%)
(2) Female	433 (19.2%)
Missing (.)	2
Provider categories	
NP/PA	220 (9.9%)
MD/DO	1951 (88.0%)
RN/nurses	47 (2.1%)
Missing (.)	36
Region	
(1) Northeast	335 (14.9%)
(2) Midwest	691 (30.7%)
(3) South	843 (37.4%)
(4) West	385 (17.1%)

2014

2,219 providers met the minimum number of eligible patients (10) for inclusion in the reliability analysis. The average number of eligible patients for providers included is 182.2 for a total of 404,406 patients. The range of numbers of patients for providers included is from 10 to 2,990. As noted above, providers with fewer than 10 eligible patient encounters during the study period were excluded.

The unit of analysis for this measure is the provider. A description of the providers studied for the 2014 calendar year is shown below:

	Total n = 2219
Provider gender	
(1) Male	1772 (79.9%)
(2) Female	447 (20.1%)
Provider categories	
NP/PA	228 (10.5%)
MD/DO	1907 (87.5%)
RN/nurses	45 (2.1%)
Missing (.)	39
Region	
(1) Northeast	334 (15.1%)
(2) Midwest	543 (24.5%)
(3) South	931 (42.0%)
(4) West	411 (18.5%)

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)*

2013

There are a total of 409,332 patients included in the temporal comparison that were treated in 2013. Patients' characteristics are provided below:

	Total
	n = 409332
Race	
(1) White	202768 (89.5%)
(2) Black	19665 (8.7%)
(3) Other	4211 (1.9%)
Missing (.)	90895
Insurance	
(0) No insurance	311 (0.6%)
(1) Private	33938 (70.5%)
(2) Medicare	13596 (28.2%)
(3) Medicaid	107 (0.2%)
(4) Other	204 (0.4%)
Missing (.)	51502
Age	
18 to <60	77884 (19.0%)
60 to <70	96129 (23.5%)
70 to <80	120506 (29.4%)
80 to 114	114813 (28.0%)
Sex	
(1) Male	228280 (55.8%)
(2) Female	180550 (44.2%)
Missing (.)	300
BMI (kg/m2)	30.6 ± 8.4
Missing	86645
Diabetes Mellitus	122242 (32.2%)
Coronary Artery Disease	241431 (66.8%)
Hypertension	312835 (86.5%)
Atrial Fibrillation/Flutter	151755 (40.4%)
Peripheral Arterial Disease	56306 (15.4%)

	Total
	n = 409332
Stroke/TIA	19269 (7.4%)
Myocardial Infarction	88104 (25.8%)

2014

There are a total of 404,406 patients included in the primary analysis (2014), whose characteristics are described below:

	Total n = 404406
Race	
(1) White	186216 (90.4%)
(2) Black	14849 (7.2%)
(3) Other	4932 (2.4%)
Missing (.)	61304
Insurance	
(0) No insurance	236 (0.5%)
(1) Private	39659 (79.2%)
(2) Medicare	9930 (19.8%)
(3) Medicaid	112 (0.2%)
(4) Other	116 (0.2%)
Missing (.)	57545
Age	
18 to <60	76319 (18.9%)
60 to <70	95585 (23.6%)
70 to <80	120341 (29.8%)
80 to 114	112161 (27.7%)
Sex	
(1) Male	224353 (55.7%)
(2) Female	178353 (44.3%)
Missing (.)	12482
BMI (kg/m2)	30.6 ± 8.2
Missing	74286
Diabetes Mellitus	118503 (30.8%)
Coronary Artery Disease	235361 (64.4%)
Hypertension	309664 (86.0%)

	Total
	n = 404406
Atrial Fibrillation/Flutter	160293 (41.5%)
Peripheral Arterial Disease	54749 (15.4%)
Stroke/TIA	21525 (7.4%)
Myocardial Infarction	84597 (24.2%)

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.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We do not currently collect any of the SDS variables examples listed above. As is noted in other sections of this testing form we do collect data on race as well as insurance type.

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: $\text{Reliability} = \text{Variance (physician-to-physician)} / [\text{Variance (physician-to-physician)} + \text{Variance (physician-specific-error)}]$, where the latter represents the within-physician estimate of our error in assessing their ‘true’ performance. Thus, this assessment of 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 at five different distributions of provider volumes: at the minimum number of quality reporting events for the measure; at the mean number of quality reporting events per physician; and at 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)

2013 – In 2013, the signal-noise ratios are shown below:

Description	Number of Patients	Signal-to-Noise Ratio
Minimum	10	0.988
25th percentile	61	0.994
50th percentile	128	0.996
75th percentile	227	0.998
Average	182	0.997

2014 – In 2014, the signal-noise ratios are shown below:

Description	Number of Patients	Signal-to-Noise Ratio
Minimum	10	0.989
25th percentile	66	0.994
50th percentile	130	0.996
75th percentile	217	0.997
Average	183	0.997

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 2013 and 2014, supporting the reproducibility of these estimates across years. At the minimum number of patient visits required (>10) the average reliability was 0.988 and 0.989 for 2013 and 2014, respectively. For providers with the median number of patient encounters, the reliability was even higher, 0.997 in both years. Given that a reliability of 0.70 is generally considered a minimum threshold for acceptability, and 0.80 is considered very good reliability, these data suggest that the measure is exceedingly good at describing true differences across physicians.

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)

Clinical Evidence: Evaluation of LVEF in patients with heart failure provides important information that is required to appropriately direct treatment. Several pharmacologic therapies have demonstrated efficacy in slowing disease progression and improving outcomes in patients with left ventricular systolic dysfunction. LVEF assessed during the initial evaluation of patients presenting with heart failure can be considered valid unless the patient has demonstrated a major change in clinical status, experienced or recovered from a clinical event, or received therapy that might have a significant effect on cardiac function. Both journal articles that are referenced in Section 1a7.9 of the evidence form supports the assessment of ejection fraction to enable appropriate guideline-directed medical treatment.

Construct validity was difficult to establish because there has not been an independent audit of these data. However, it is important to note that an independent audit would merely involve an abstractor reviewing the same medical record from which PINNACLE directly abstracts its data and, given the identical source of the data, any error observed would either be due to the auditor incorrectly abstracting the data from the EHR or PINNACLE incorrectly mapping the data elements from the EHR. To address the latter, we conduct detailed analyses to insure that this does not happen and quarantine (i.e. not report) data that fails our additional Data Quality Review process. Validity of measure data elements in PINNACLE is routinely evaluated on a quarterly basis as part of the standard data extraction and analytic data set creation process. First, all relevant data elements are reviewed at the record level to ensure that individual data values are valid; any invalid values are set to missing. Next, the *distribution* of each data element is reviewed, aggregating both across practices and across calendar quarters within each practice, to identify outliers, suspicious patterns and/or systematic changes in the prevalence of

the data element that may suggest data mapping errors or unanticipated changes in definitions, coding consistency, data completeness, etc. Identification of suspicious data includes both statistical criteria, using quality control charts with rigorous definitions of “out of control” rates, and manual clinical review of each distribution for plausibility. Records that are flagged as suspicious by these criteria are quarantined and excluded from analysis and reporting. In 2013 the rate of records not passing the quality evaluation was 3.1% and in 2014 it was 8.7%. Feedback reports are generated to facilitate investigation of data issues at the practice level to verify accuracy of abstraction and to remap elements whose definitions or recording have changed.

Content validity for this measure was assessed by expert work group members during the development process. Additional input on the content validity of draft measures was established through a 30-day public comment period and concurrent formal peer review process. 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, FACP (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 (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); Paul D. Rockswold, MD, MPH (family medicine); Ileana L. Piña, MD, FACC (cardiology, heart failure); 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).

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 and one at 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

Forty two (42) committee members completed the survey and provided a mean importance rating of 4.24, with 85.7% agreeing with the use of the measure for quality assessment.

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

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

N = 42; Mean rating = 4.24 and 85.7% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

Frequency Distribution of Ratings

1 - <2> (Strongly Disagree)

2 - <3>

3 - <1> (Neither Agree nor Disagree)

4 - <13>

5 - <23> (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 high face validity by both its clinical importance and by 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)

Since not all patients with heart failure, in rare circumstances, might not meet the guideline recommendations for EF evaluation, exclusions in this measure are intended to remove patients for whom an EF assessment might not be appropriate (e.g. patient refusal). We divide these into two categories: Exclusions and Exceptions. Exclusions arise when patients who are included in the initial patient or eligible population for the measure set do not meet the denominator criteria specific to the intervention required by the numerator. There are no known exclusions for this assessment, which are usually derived from evidence-based guidelines. Exceptions are not absolute, and are based on clinical judgment and individual patient characteristics, such as refusal to participate in the test.

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)

There are no exclusions and there were no exceptions reported for this measure in any patients.

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)

All patients are eligible for this measure and all providers caring for HF patients should be able to have a performance rate calculated on their entire population.

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 (clinical factors or sociodemographic 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)

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

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

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)

We examined variation in provider performance on this measure based on sex, age, race and a number of other patient factors to identify variations. The findings are represented for 2013 and 2014 respectively.

2013

Label	# of providers	# of patients	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
Male	2250	228280	0.00%	47.6%	69.3%	95.5%	100%	47.8%	31.8%
Female	2250	180550	0.00%	41.0%	66.2%	94.4%	100%	53.4%	32.8%
Age: <60	2236	77884	0.00%	42.0%	66.1%	94.4%	100%	52.4%	33.3%
Age: 60 -< 70	2246	96129	0.00%	46.3%	68.4%	96.4%	100%	50.1%	33.0%
Age: 70 -< 80	2253	120506	0.00%	46.8%	69.4%	96.9%	100%	50.1%	32.8%
Age: >= 80	2243	114813	0.00%	42.9%	67.3%	96.0%	100%	53.1%	33.4%
Insurance: None	69	311	0.00%	25.0%	69.1%	100%	100%	75.0%	42.7%
Insurance: Private	596	33938	0.00%	33.3%	64.4%	97.1%	100%	63.8%	35.7%
Insurance: Medicaid	382	13596	0.00%	28.6%	60.0%	90.2%	100%	61.6%	35.6%
Insurance: Medicare	23	107	0.00%	0.00%	47.6%	91.7%	100%	91.7%	44.2%
Insurance: Other	76	204	0.00%	50.0%	76.3%	100%	100%	50.0%	36.3%
Race: White	1479	202768	0.00%	50.8%	71.2%	96.9%	100%	46.1%	31.4%

Label	# of providers	# of patients	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
Race: Black	1267	19665	0.00%	48.6%	70.5%	100%	100%	51.4%	36.3%
Race: Other	842	4211	0.00%	33.3%	67.2%	100%	100%	66.7%	40.6%

2014

label	# of providers	# of patients	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
Male	2214	224353	0.00%	56.6%	73.8%	96.6%	100%	39.9%	29.5%
Female	2214	178353	0.00%	51.2%	71.1%	95.6%	100%	44.4%	30.7%
Age: <60	2203	76319	0.00%	50.0%	70.1%	95.7%	100%	45.7%	31.3%
Age: 60 -< 70	2213	95585	0.00%	54.5%	72.6%	97.3%	100%	42.8%	31.1%
Age: 70 -< 80	2216	120341	0.00%	58.4%	74.4%	97.9%	100%	39.6%	30.3%
Age: >= 80	2217	112161	0.00%	52.6%	72.2%	97.1%	100%	44.5%	31.0%
Insurance: None	53	236	0.00%	0.00%	67.2%	100%	100%	100%	45.5%
Insurance: Private	586	39659	0.00%	54.2%	73.7%	99.1%	100%	45.0%	30.5%
Insurance: Medicaid	342	9930	0.00%	50.0%	69.3%	100%	100%	50.0%	33.7%
Insurance: Medicare	24	112	0.00%	20.0%	63.1%	100%	100%	80.0%	40.7%
Insurance: Other	55	116	0.00%	0.00%	65.4%	100%	100%	100%	42.8%
Race: White	1248	186216	0.00%	58.0%	74.2%	97.5%	100%	39.4%	30.4%
Race: Black	1070	14849	0.00%	50.0%	72.1%	100%	100%	50.0%	35.9%
Race: Other	761	4932	0.00%	50.0%	71.7%	100%	100%	50.0%	38.4%

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)

Provided below is the testing that identified the differences in performance measure scores for 2013 and 2014.

2013

Overall mean performance on this measure is 67.8%, with a standard deviation of 32.0%. The minimum score equals 0.00%, while the maximum score equals 100.00%. The interquartile score is equal to 49.6%.

2,254 providers were measured, and the patient study sample equals 409,332. 55.8% of the sample is male. 89.5% of the sample is white, 8.7% is black, and 1.9% identified as "other." The sample reached across all US regions, with 14.9% of providers in the Northeast, 30.7% of providers in the Midwest, 37.4% of providers in the South, and 17.1% of providers in the West.

# of providers	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
2254	0.00%	45.2%	67.8%	94.8%	100%	49.6%	32.0%

	Mean
Decile 1	4.0%
Decile 2	23.5%
Decile 3	44.5%
Decile 4	62.3%
Decile 5	75.5%
Decile 6	84.9%
Decile 7	90.6%
Decile 8	94.7%
Decile 9	97.7%
Decile 10	99.7%

2014

Overall mean performance on this measure is 72.5%, with a standard deviation of 29.9%. The minimum score equals 0.00%, while the maximum score equals 100.00%. The interquartile score is equal to 41.6%.

2,219 providers were measured, and the patient study sample equals 404,406. 55.7% of the sample is male. 90.4% of the sample is white, 7.2% is black, and 2.4% identified as "other." The sample reached across all US regions, with 15.1% of providers in the Northeast, 24.5% of providers in the Midwest, 42.0% of providers in the South, and 18.5 % of providers in the West.

# of providers	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
2219	0.00%	54.3%	72.5%	95.9%	100%	41.6%	29.9%

	Mean
Decile 1	7.2%
Decile 2	34.7%
Decile 3	54.4%
Decile 4	69.9%
Decile 5	82.1%
Decile 6	89.4%
Decile 7	93.1%
Decile 8	95.9%
Decile 9	98.0%
Decile 10	99.7%

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?*)

2013: A moderate to large amount of variability was noted among providers. The performance-met rate range was 0-100% with the inter-quartile range being 45.2% to 94.8%. This yielded a Median Rate Ratio of 2.13 (2.07, 2.18). The Median Rate Ratio measures the variation across providers for statistically 'identical' patients and suggests that a patient presenting to 1 provider, as opposed to another, would, on average, be 2.3-times more likely to have their EF assessed.

2014: A moderate, but slightly lower, amount of variability was noted among providers. The performance-met rate range was 0-100% with the inter-quartile range being 54.3% to 95.9%. This yielded a Median Rate Ratio of 1.92 (1.88, 1.97).

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 item is directed to measures that are risk-adjusted (with or without SDS factors) **OR** 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). **Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.**

2b6.1. Describe the method of testing conducted to compare 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*)

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 the differences in 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

2b7.1. 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 LVEF of the patient, but did not record them in the electronic health record, from which the data in PINNACLE is directly abstracted. If there are missing fields, we believe that these will be rapidly corrected on 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 frequency or distribution of missing data. For this measure, missing data is reported as a quality 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*)

We do not believe any biases are introduced in the assessing of individual physician performance and continued endorsement of this measure would lead to improved care.