

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

Measure Number (if previously endorsed): 1524

Measure Title: AF: Assessment of Thromboembolic Risk Factors (CHADS₂)

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 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 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.

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 included all patients with atrial fibrillation in the PINNACLE Registry occurring 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. 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. In addition, practices where 0 patient encounters across all physicians met the measure, we assumed that there was a residual error in coding and excluded those practices. 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? 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)*

912 providers met the minimum number of eligible patients (10) for inclusion in the primary analysis (2012 data). The average number of eligible patients for providers included is 243.5. The range of number of patients for providers included is from 1,885 to 10. A description of the providers is shown below:

	Total
	n = 912
Provider gender	
(1) Male	716 (78.8%)
(2) Female	193 (21.2%)
Missing (.)	3
Provider categories	
NP/PA	100 (11.1%)
MD/DO	767 (85.2%)
RN/nurses	33 (3.7%)
Missing (.)	12
Region	
(1) Northeast	151 (16.6%)
(2) Midwest	255 (28.0%)
(3) South	317 (34.8%)
(4) West	189 (20.7%)

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 are a total of 222063 patient encounters included in the primary analysis (2012). See Table below provides details on these patients' demographic and clinical characteristics.

	Total
	n = 222063
Race	
(1) White	145446 (93.8%)
(2) Black	7660 (4.9%)
(3) Other	1926 (1.2%)
Missing (.)	67031

Insurance	
(0) No insurance	13218 (6.4%)
(1) Private	123434 (59.8%)
(2) Medicare	66025 (32.0%)
(3) Medicaid	2348 (1.1%)
(4) Other	1278 (0.6%)
Missing (.)	15760
Age	
18 to <60	32312 (14.6%)
60 to <70	50550 (22.8%)
70 to <80	69007 (31.1%)
80 to 112	70194 (31.6%)
Sex	
(1) Male	125556 (56.5%)
(2) Female	96489 (43.5%)
Missing (.)	18
BMI	29.7 ± 6.8
Missing	51041
Diabetes	63705 (28.7%)
CAD	140797 (63.4%)
Hypertension	195878 (88.2%)
AFib	222063 (100.0%)
HF	86545 (39.0%)
PAD	42596 (19.2%)
Prior Stroke/TIA	54804 (24.7%)
MI history	61456 (27.7%)

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: $\text{Reliability} = \text{Variance (physician-to-physician)} / [\text{Variance (physician-to-physician)} + \text{Variance (physician-specific-error)}]$, where the latter represents the within-physician estimate of our error in assessing their 'true' performance. 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 volume levels of physician visits: at the minimum number of quality reporting events for the measure; at 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	0.990
25th percentile	91	0.996
50th percentile	159	0.997
75th percentile	247	0.998
Average	196	0.998

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

Description	Number of Patients	Signal-to-Noise Ratio
Minimum	10	0.993
25th percentile	133	0.996
50th percentile	216	0.997
75th percentile	309	0.998
Average	244	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 2011 and 2012, supporting the reproducibility of these estimates across years. At the minimum number of patient visits required (>10) the average reliability was 0.990 and 0.993 for 2011 and 2012, 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)

There is a strong clinical foundation for assessing the thromboembolic risk of patients with atrial fibrillation because treatment with systemic anti-coagulation is strongly endorsed by clinical trials in patients at higher risk for thromboembolism. While the risk stratification score (e.g. CHADS₂ or CHADS-Vasc) may vary, guidelines and numerous clinical and observational studies demonstrate an increasing benefit:risk ratio with warfarin, or a novel anti-coagulant, in patients with higher thromboembolic risk. Thus a cornerstone of treating atrial fibrillation patients is knowing their thromboembolic risk.

To further establish the face validity of the measure, the ACC and AHA solicited expert input on this performance measure. After the measure was fully specified, members of three 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

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 = 17; Mean rating = 4.53 and 94.1% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality. The distribution of responses is provided below:

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 1

3 - 0 (Neither Agree nor Disagree)

4 - 5

5 – 11-(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. There were 17 committee members who completed the survey and provided a mean rating of 4.53 (94.1% either agreed or strongly agreed with the importance and value of this measure), showing strong support for this performance measure.

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 this measure is a first step in determining whether anticoagulation has been prescribed for eligible patients with non-valvular Atrial Fibrillation (AF), exclusions in this measure are intended to remove patients whose AF is caused by valvular disease or for whom chronic anticoagulation would be contraindicated (transient or reversible causes of AF). 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. Exclusions are absolute and apply to all patients and therefore are not part

of clinical judgment within a measure. Specific exclusions should be derived from evidence-based guidelines.

Exclusions in this measure:

- Patients with mitral stenosis or prosthetic heart valves
- Patients with transient or reversible causes of AF (eg, pneumonia, hyperthyroidism, pregnancy, cardiac surgery)

In the context of physician performance measurement, exceptions are the mechanism used to remove patients from the denominator of a performance measure when a patient does not receive a therapy or service AND that therapy or service would not be appropriate due to specific reasons for which the patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment and individual patient characteristics. In this case, assessing thromboembolic risk, there are no conceivable patient, or system reasons for not documenting risk factors and rare medical reasons not to do so. In 2012, 98.9% reported no exclusions and among the providers who did report exclusions, the mean rate was 2.1%, with a range of 0.3-6.9% of their patients.

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)*

2011: 98.4%(n=694) of the providers do not have exclusions. Among the 11 providers who do have exceptions, the exclusion rate ranges from 0.2% to 9.8%, mean is 3.1%.

2012: 98.9% (n=902) of the providers do not have exclusions. Among the providers who do have exceptions, the exclusion rate ranges from 0.3% to 6.9%, mean is 2.1%.

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)*

The overall frequency of exceptions is very low. The distribution of exceptions across physicians is narrow, indicating that the occurrence of an exception is very low and not likely to bias performance results. We believe all of the exclusions must stay in the measure based upon the clinical imperative for anti-coagulation in these settings (e.g. mitral stenosis or valvular atrial fibrillation) or the lack of benefit in those with transient and reversible atrial fibrillation) and the exceptions should also remain in the measure because they represent valid reasons that patients would not be expected to be assessed for thromboembolic risk factors and allow physicians to customize their care to weigh the risks and benefits for an individual patient.

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

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 and provide additional information about potential disparities based on sex, age, race and a number of other patient factors. The full testing report with information on all patient characteristics is available in Appendix A-1 and summarized below.

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

# of providers	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
705	0.00%	0.39%	22.8%	32.0%	100%	31.6%	33.2%

2012

# of providers	Minimum	Lower Quartile	Mean	Upper Quartile	Maximum	Quartile Range	Std Dev
912	0.00%	0.00%	20.5 %	25.4%	100%	25.4%	30.6%

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 patients with atrial fibrillation, ranging from providers who assessed thromboembolic risk in none of their patients to others who assessed it in every patient. This was observed in both 2011 and 2012, with the latter including even more providers. 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 thromboembolic risk. 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 risk assessed.

2011: A large amount of variability was noted among providers. The performance-met rate range was 0-100% with the inter-quartile range being 0.4% to 32%. This yielded a Median Rate Ratio of 7.9 (7.0, 9.0). The Median Rate Ratio measures the variation between clusters by comparing 2 persons from two randomly chosen different clusters. A MRR of 7.9 is an enormous amount of variation, meaning that the 'same' patient seeing 2 different providers would be almost 8-fold more likely to have their risk assessed by 1 provider vs. another.

2012: A large amount of variability was noted among providers. The performance-met rate range was 0-100% with the inter-quartile range being 0% to 25%. This yielded a Median Rate Ratio of 9.13 (8.16, 10.34). The Median Rate Ratio measures the variation between clusters by comparing 2 persons from two randomly chosen different clusters. A MRR of 9.13 is an enormous amount of variation, meaning that the 'same' patient seeing 2 different providers would be more than 9-fold more likely to have their risk assessed by 1 provider vs. another.

Given the clinical importance of explicitly defining patients' thromboembolic risks, and the wide variation observed across precision with a very high signal-noise ratio, we believe that this is an excellent measure for detecting providers with better or worse performance.

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*)

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

2b7.1. 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, missing values are interpreted as 'No' for most of variables. For example, Thromboembolic Risk Factors Assessed: missing - not assessed; 1 - Yes (All risk factors assessed); 2 - No - Medical Reason; 3 - No - Patient Reason; 4 - No - System Reason. It's challenging to distinguish real missing vs 'No'. However, we do think it's reasonable to assume that data were not collected (missing) if all records from a practice are missing. For 2011 data, we identified 44 such practices for Thromboembolic Risk Factors Assessed. For 2012 data, we identified 33 such practices for Thromboembolic Risk Factors Assessed. These practices are excluded from the analysis. While this is a logical approach for handling missing data at the practice level, there are likely missing data (i.e. poor documentation) across providers. However, this is not a threat to the validity of the measure, as the ability to more clearly document thromboembolic risk is clearly under the locus of control of the provider. We believe that if this is an endorsed measure and is used to assess the quality of atrial fibrillation care, that there will be increasing pressure on providers both to assess thromboembolic risk and to clearly document this risk – the major goals of quality improvement.

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 to assess thromboembolic risk.

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 and that providers with a very low rate of assessing thromboembolic risk either failed to assess patients' risks, or failed to document that they did so. Either way, we do not believe that any biases are introduced in assessing individual physician performance and endorsement of this measure would lead to improved care.