

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

Measure Number (if previously endorsed): Not applicable.

Measure Title: Cardiac Tamponade and/or Pericardiocentesis Following Atrial Fibrillation Ablation.

Date of Submission: 12/23/2013

Type of Measure: Outcome.

☐ 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](#).

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 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:
☐ 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).

This project utilized a large data warehouse compiled by Inovalon, Inc., the Medical Outcomes Research for Effectiveness and Economics Registry (MORE2 Registry®). This warehouse of healthcare data contains information derived from more than 7.7 billion medical events generated by over 99.6 million unique and de-identified individuals nationwide. Encompassing more than 3.2 billion member-months of data, the registry goes beyond claims data to include information about demographics, enrollment, diagnoses, procedures, pharmacy and laboratory results and represents a significant mix of commercial, Medicare Advantage, and managed Medicaid plan members. The data reside on an IBM-Netezza data warehouse server.

Age Range	MORE ²		Insured U.S. Population*	
	Unique Persons	Percent	Unique Persons	Percent
0-9	17,865,773	16.9%	37,137,858	14.6%
10-19	17,085,891	16.2%	36,289,407	14.3%
20-29	14,092,936	13.3%	27,954,341	11.0%
30-39	12,039,891	11.4%	29,073,246	11.5%
40-49	11,938,741	11.3%	30,756,671	12.1%
50-59	11,125,355	10.5%	35,820,293	14.1%
60-69	9,738,459	9.2%	28,304,711	11.2%
70-79	6,560,098	6.2%	17,197,678	6.8%
80+	4,829,224	4.6%	10,983,438	4.3%
Unknown	401,901	0.4%	-	0.0%
Total	105,678,269		253,517,642	

* Source: 2011 Medical Expenditure Panel Survey (MEPS)

1.3. What are the dates of the data used in testing? 2007-2012

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

Measure Specified to Measure Performance of: (must be consistent with levels entered in item S.26)	Measure Tested at Level of:
☒ 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)

The database includes 666,617 physicians.

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)

The database includes approximately 99.6 million unique and de-identified individuals' records.

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.

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)

The statistical reliability scores are based on an analysis of the relative value of variation in measure scores due to signal (i.e., variation between measured entities) versus noise (i.e., variation within measured entities) using a "signal to noise test" statistic based on a beta-binomial model described in a paper by John L. Adams for the National Committee for Quality Assurance ("The Reliability of Provider

Profiling: A Tutorial", 2009, RAND Corporation). Reliability scores range from 0.0 to 1.00. Generally, a reliability score of .70 or higher is considered good (the higher the better).

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)

HRS-12: Three-year Period Reliability Scores by by Provider Type						
	Physician 3-Year Avg.			Facility 3-Year Avg.		
	Avg. Reliability Score			Avg. Reliability Score		
	Minimum Procedures			Minimum Procedures		
<u>Years</u>	<u>2</u>	<u>10</u>	<u>20</u>	<u>2</u>	<u>10</u>	<u>20</u>
2007-2009	94.3%	84.4%	78.9%	91.9%	80.8%	69.3%
2008-2010	93.2%	83.2%	78.9%	90.6%	76.9%	65.0%
2009-2011	93.3%	84.6%	75.5%	91.6%	78.9%	71.5%
2010-2012	93.8%	83.4%	68.4%	90.8%	75.1%	61.1%
	Count of Physicians			Count of Facilities		
	Minimum Procedures			Minimum Procedures		
<u>Years</u>	<u>2</u>	<u>10</u>	<u>20</u>	<u>2</u>	<u>10</u>	<u>20</u>
2007-2009	2,535	388	127	1,996	568	262
2008-2010	2,949	445	144	2,068	581	292
2009-2011	2,734	399	105	1,757	486	240
2010-2012	2,296	299	68	1,519	395	175
	Estimate of Mu			Estimate of Mu		
	Minimum Procedures			Minimum Procedures		
<u>Years</u>	<u>2</u>	<u>10</u>	<u>20</u>	<u>2</u>	<u>10</u>	<u>20</u>
2007-2009	1.43%	1.17%	1.14%	1.38%	1.33%	1.37%
2008-2010	1.64%	1.08%	0.80%	1.46%	1.33%	1.39%
2009-2011	1.42%	0.95%	0.98%	1.38%	1.34%	1.40%
2010-2012	1.30%	1.13%	1.20%	1.36%	1.38%	1.49%
	Estimate of Alpha			Estimate of Alpha		
	Minimum Procedures			Minimum Procedures		
<u>Years</u>	<u>2</u>	<u>10</u>	<u>20</u>	<u>2</u>	<u>10</u>	<u>20</u>
2007-2009	0.3114	0.7870	0.7813	0.6733	1.1764	2.1955
2008-2010	0.3627	1.4171	1.9254	1.0540	2.4620	3.8363
2009-2011	0.7685	3.7448	1.9316	0.7881	1.3692	1.4794
2010-2012	1.9127	3.8970	11.6531	3.3033	8.9849	12.6406
	Estimate of Beta			Estimate of Beta		
	Minimum Procedures			Minimum Procedures		
<u>Years</u>	<u>2</u>	<u>10</u>	<u>20</u>	<u>2</u>	<u>10</u>	<u>20</u>
2007-2009	21.49	66.21	67.58	47.94	87.18	157.53
2008-2010	21.71	129.87	238.17	71.36	182.48	271.88
2009-2011	53.51	390.90	194.49	56.45	101.15	104.06
2010-2012	145.11	340.30	956.84	238.88	642.04	837.22

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

The results indicate that the measure is reliable at physician and facility level of analysis.

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)

Validity testing of the computed measure scores was completed to assure measure scores do not vary by type of data (i.e., provider level data vs. hospital level data) and to provide evidence that the measure can be used for the purpose of differentiating quality of care across providers and facilities. Measure scores were also compared across age groups, gender groups and by procedure type, with resulting measure rates consistent with expected outcomes and the literature.

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

Distribution of HRS-12 for Physicians											
Minimum Denominator	Time Period	N	Percentile								Maximum
			25th	50th	75th	80th	85th	90th	95th	99th	
2	2007-2009	2,535	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.69%	50.00%	100.00%
	2008-2010	2,949	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	11.11%	50.00%	100.00%
	2009-2011	2,734	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	10.00%	50.00%	100.00%
	2010-2012	2,296	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7.69%	50.00%	100.00%
10	2007-2009	388	0.00%	0.00%	0.00%	0.00%	3.13%	5.56%	8.33%	14.29%	18.18%
	2008-2010	445	0.00%	0.00%	0.00%	0.00%	3.13%	5.56%	8.33%	14.29%	18.18%
	2009-2011	399	0.00%	0.00%	0.00%	0.00%	1.79%	5.00%	7.14%	10.00%	11.76%
	2010-2012	299	0.00%	0.00%	0.00%	0.00%	2.94%	5.88%	8.33%	18.18%	20.00%
20	2007-2009	127	0.00%	0.00%	1.32%	2.52%	3.23%	4.55%	6.90%	12.00%	13.64%
	2008-2010	144	0.00%	0.00%	0.00%	1.67%	3.03%	3.85%	4.35%	6.06%	9.09%
	2009-2011	105	0.00%	0.00%	1.01%	2.50%	3.03%	3.57%	4.35%	6.82%	8.70%
	2010-2012	68	0.00%	0.00%	2.60%	3.23%	3.70%	4.76%	5.71%	8.33%	8.33%

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?)

In reviewing rates by ablation type, the data validate that catheter procedure is by far the most common

type of ablation completed and that rates of complications are lower, consistent with clinical knowledge. The rates of complications are far higher for less frequently performed procedures (extensive and limited approaches), also consistent with expectations. Similarly, complication rates are higher in older patients, consistent with clinical expectations of greater risk of complications in older patients. Complication rates are also higher in females as compared to males, again supported by the literature.

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*)

Not applicable.

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*)

Not applicable.

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*)

Not applicable.

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 2 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.

Not applicable.

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*)
Age and gender were selected as part of the stratification model as research indicates that these factors contribute to the risk of this complication.

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

There were no statistical analyses used to select the risk factors. The two factors were determined by a Committee of physician experts who affirmed that these factors contribute to the risk of complication.

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*)

There was no statistical analysis used to develop and validate the adequacy of the stratification approach. The two factors were determined by a Committee of physician experts who affirmed that these factors contribute to the risk of complication.

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*):

Not applicable.

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

Not applicable.

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

Not applicable.

2b4.9. Results of Risk Stratification Analysis:

The results showed that the rate of complication was 1.25 times higher for women (1.28%) compared to men (1.02%). These results are consistent with the literature.

The results demonstrated that the rate of complication was slightly higher for patients 65 years of age and old (1.43%) compared to patients 18 to 64 years of age (1.23%).

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*)

Rather than controlling for differences in patient characteristics, the measure stratifies results by two patient variables identified by our committee of physician experts and corroborated by contemporary peer-reviewed scientific literature—age and gender. Measure testing confirmed that AF ablation complication rates are higher in females and in patients ≥ 65 years of age, supporting the recommendation to stratify results accordingly.

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)

Validity testing of the computed measure scores was completed to provide evidence that the measure can be used for the purpose of differentiating quality of care across providers and facilities. Measure scores were also compared across age groups, gender groups and by procedure type, with resulting measure rates consistent with expected outcomes and the literature. For example, in reviewing rates by ablation type, the data validate that catheter procedure is by far the most common type of ablation completed and that rates of complications are lower, consistent with clinical knowledge. The rates of complications are far higher for less frequently performed procedures (extensive and limited approaches), also consistent with expectations. Similarly, we see that rates are higher in older patients, consistent with clinical expectations of greater risk of complications in older patients. Complication rates are also higher in females as compared to males, again supported by the literature.

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)

Years	Average HRS-12								
	Minimum Number of Procedures by Physician								
	1	2	3	5	10	15	20	30	60
2007-2009	1.59%	1.63%	1.38%	1.28%	1.20%	1.19%	1.21%	0.79%	1.31%
2008-2010	1.85%	1.98%	1.67%	1.45%	1.18%	1.00%	0.79%	0.80%	0.61%
2009-2011	1.71%	1.69%	1.53%	1.37%	0.92%	1.02%	0.94%	1.33%	0.72%
2010-2012	1.57%	1.50%	1.28%	1.20%	1.17%	1.05%	1.25%	0.93%	0.95%
Years	HRS-12 Standard Error								
	Minimum Number of Procedures by Physician								
	1	2	3	5	10	15	20	30	60
2007-2009	0.15%	0.17%	0.15%	0.14%	0.16%	0.19%	0.23%	0.18%	0.44%
2008-2010	0.15%	0.17%	0.15%	0.15%	0.14%	0.15%	0.14%	0.18%	0.25%
2009-2011	0.15%	0.15%	0.15%	0.15%	0.12%	0.15%	0.18%	0.27%	0.37%
2010-2012	0.15%	0.15%	0.14%	0.15%	0.18%	0.19%	0.26%	0.24%	0.41%

Please also see Appendix 1.

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?)

Please see Appendix 1.

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*)

Not applicable.

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*)

Not applicable.

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*)

Not applicable.

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*)

The threat of bias due to missing data was addressed by calculating performance results based on varying numbers of minimum procedures.

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*)

Results demonstrate good validity at a minimum of 10 procedures as compared to minimum of 20 procedures. The results also show better reliability at the physician level as compared to the facility level, which is consistent with expectations, and shows the complication rates are more dependent on the physician performing the procedure than the hospital facility at which the procedure is performed.

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*)

Performance results were consistent with expectations, demonstrating that complication rates are more dependent on the physician performing the procedure than the hospital facility at which the procedure is performed.