

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

Measure Number (if previously endorsed): 2411

Measure Title: PCI: Comprehensive Documentation of Indications for PCI

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 elective PCI's in the CathPCI Registry performed during the one-year study period.

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

The primary analysis included all elective PCI's in the CATHPCI Registry from 1/1/2012 thru 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 type="checkbox"/> individual clinician	<input type="checkbox"/> individual clinician
<input type="checkbox"/> group/practice	<input type="checkbox"/> group/practice
<input checked="" type="checkbox"/> hospital/facility/agency	<input checked="" 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)

1309 hospitals across the U.S. were included in the primary analysis. The table below summarizes the distribution by hospital PCI volume, location, type and region. Hospitals are the primary level of measurement and analysis.

Description of sites in 2012

	Total n = 1309
AUC	
PCI Volume in 2012	503.3 ± 424.8
Hospital Location	
RURAL	240 (18.3%)
SUBURBAN	460 (35.1%)
URBAN	609 (46.5%)
Participant Type	
GOVERNMENT	19 (1.5%)
PRIVATE/COMMUNITY	1178 (90.0%)
UNIVERSITY	112 (8.6%)
Teaching Hospital	502 (38.3%)
Public Hospital	483 (36.9%)
Census Region	
MIDWEST REGION	377 (28.8%)
NORTHEAST REGION	172 (13.1%)
SOUTH REGION	507 (38.8%)
WEST REGION	252 (19.3%)
Missing	1

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

There were 135,298 patients undergoing elective PCI that were included in the primary analysis. Of these, 94,735 were male (70%) and 40,563 were female (30%). 118,007 were Caucasian (87.22%), 10,421 were African American (7.70%) and 6,870 were other races (5.08%). A complete description of patients, stratified by whether or not the indications for PCI were clearly described (i.e. the appropriateness of the PCI procedure was ‘mappable’ or not), is provided in the table below.

AUC			
	Total		
	n = 135298	Not Mappable to AUC n = 49063	Mappable to AUC n = 86235
History			
Age	66.5 ± 10.9	66.5 ± 11.1	66.5 ± 10.9
Sex			
Male	94735 (70.0%)	34906 (71.1%)	59829 (69.4%)
Female	40563 (30.0%)	14157 (28.9%)	26406 (30.6%)
IABP	771 (0.6%)	313 (0.6%)	458 (0.5%)
Missing (.)	36	12	24
Current/Recent Smoker (w/in 1 year)	27104 (20.0%)	10399 (21.2%)	16705 (19.4%)
Missing (.)	91	46	45
Hypertension	117498 (86.9%)	42403 (86.5%)	75095 (87.1%)
Missing (.)	40	18	22
Dyslipidemia	115183 (85.2%)	41784 (85.2%)	73399 (85.2%)
Missing (.)	105	44	61
Family History of Premature CAD	33607 (24.9%)	12125 (24.7%)	21482 (24.9%)
Missing (.)	62	33	29
Prior MI	43233 (32.0%)	18828 (38.4%)	24405 (28.3%)
Missing (.)	34	22	12
Prior Heart Failure	18974 (14.0%)	7435 (15.2%)	11539 (13.4%)
Missing (.)	74	41	33
Prior Valve Surgery/Procedure	2505 (1.9%)	1081 (2.2%)	1424 (1.7%)
Missing (.)	77	36	41
Prior PCI	63566 (47.0%)	26145 (53.3%)	37421 (43.4%)
Missing (.)	29	19	10
Prior CABG	26026 (19.2%)	13587 (27.7%)	12439 (14.4%)
Missing (.)	16	16	
Currently on Dialysis	3957 (2.9%)	1374 (2.8%)	2583 (3.0%)
Missing (.)	152	63	89
Cerebrovascular Disease	18068 (13.4%)	6652 (13.6%)	11416 (13.2%)
Missing (.)	63	30	33
Peripheral Arterial Disease	19454 (14.4%)	7148 (14.6%)	12306 (14.3%)
Missing (.)	72	35	37
Chronic Lung Disease	19754 (14.6%)	7115 (14.5%)	12639 (14.7%)
Missing (.)	66	31	35
Diabetes Mellitus	54312 (40.2%)	19668 (40.1%)	34644 (40.2%)
Missing (.)	77	36	41
Cath Lab Visit			
PCI Indication			
Staged PCI	29487 (21.8%)	20308 (41.4%)	9179 (10.6%)
Other	105754 (78.2%)	28738 (58.6%)	77016 (89.4%)
Missing (.)	57	17	40

AUC			
	Total	Not Mappable to AUC n = 49063	Mappable to AUC n = 86235
	n = 135298		
CAD Presentation			
No symptom, no angina	36791 (27.2%)	18006 (36.7%)	18785 (21.8%)
Symptom unlikely to be ischemic	13069 (9.7%)	4266 (8.7%)	8803 (10.2%)
Stable angina	85438 (63.1%)	26791 (54.6%)	58647 (68.0%)
Anginal Classification w/in 2 Weeks			
No symptoms	40621 (30.1%)	19208 (39.3%)	21413 (24.8%)
CCS I	14016 (10.4%)	5477 (11.2%)	8539 (9.9%)
CCS II	50144 (37.1%)	18197 (37.2%)	31947 (37.0%)
CCS III	26204 (19.4%)	4736 (9.7%)	21468 (24.9%)
CCS IV	4116 (3.0%)	1248 (2.6%)	2868 (3.3%)
Missing (.)	197	197	
Anti-Anginal Medication w/in 2 Weeks	103853 (76.8%)	38304 (78.1%)	65549 (76.0%)
Missing (.)	46	37	9
Heart Failure w/in 2 Weeks	13711 (10.1%)	4890 (10.0%)	8821 (10.2%)
Missing (.)	58	20	38
Cardiomyopathy or Left Ventricular Systolic Dysfunction	18071 (13.4%)	6594 (13.4%)	11477 (13.3%)
Missing (.)	26	11	15
Pre-operative Evaluation Before Non-Cardiac Surgery	7615 (5.6%)	2081 (4.2%)	5534 (6.4%)
Missing (.)	31	19	12
Cardiogenic Shock w/in 24 Hours	487 (0.4%)	198 (0.4%)	289 (0.3%)
Missing (.)	13	2	11
Cardiac Arrest w/in 24 Hours	614 (0.5%)	263 (0.5%)	351 (0.4%)
Missing (.)	32	15	17
Pre-PCI Left Ventricular Ejection Fraction	53.5 ± 12.3	52.4 ± 12.6	54.1 ± 12.2
Missing	30945	13843	17102
Procedure Information			
Contrast Volume	186.0 ± 90.6	179.3 ± 91.5	189.9 ± 89.8
Missing	333	131	202
Fluoroscopy Time	15.6 ± 12.8	15.2 ± 12.9	15.8 ± 12.8
Missing	1704	636	1068
Outcomes			
Discharge Status			
Alive	134789 (99.6%)	48854 (99.6%)	85935 (99.7%)
Deceased	509 (0.4%)	209 (0.4%)	300 (0.3%)
Primary Cause of Death			
Cardiac	328 (64.6%)	132 (63.2%)	196 (65.6%)
Neurologic	33 (6.5%)	17 (8.1%)	16 (5.4%)
Renal	10 (2.0%)	2 (1.0%)	8 (2.7%)
Vascular	8 (1.6%)	1 (0.5%)	7 (2.3%)
Infection	19 (3.7%)	8 (3.8%)	11 (3.7%)
Valvular	18 (3.5%)	8 (3.8%)	10 (3.3%)
Pulmonary	46 (9.1%)	17 (8.1%)	29 (9.7%)
Unknown	18 (3.5%)	12 (5.7%)	6 (2.0%)
Other	28 (5.5%)	12 (5.7%)	16 (5.4%)
Missing (.)	134790	48854	85936

AUC			
	Total	Not Mappable to AUC n = 49063	Mappable to AUC n = 86235
	n = 135298		
Myocardial Infarction (Biomarker Positive) Missing (.)	2644 (2.0%) 34	929 (1.9%) 16	1715 (2.0%) 18
Cardiogenic Shock Missing (.)	389 (0.3%) 35	146 (0.3%) 17	243 (0.3%) 18
Heart Failure Missing (.)	483 (0.4%) 37	183 (0.4%) 18	300 (0.3%) 19
CVA/Stroke Missing (.)	152 (0.1%) 38	57 (0.1%) 18	95 (0.1%) 20
Other Vascular Complications Requiring Treatment Missing (.)	499 (0.4%) 41	182 (0.4%) 18	317 (0.4%) 23
RBC/Whole Blood Transfusion Missing (.)	1640 (1.2%) 38	656 (1.3%) 17	984 (1.1%) 21

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-abtractor 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 hospital performance. Reliability at the level of the specific hospital is given by: Reliability = Variance (hospital-to-hospital) / [Variance (hospital-to-hospital) + Variance (hospital-specific-error)]. A reliability of zero implies that all the variability in a measure is attributable to

measurement error. A reliability of 1.0 implies that all the variability is attributable to real differences in hospital performance.

Reliability testing was estimated by using a beta-binomial model. The beta-binomial model assumes the hospital's performance is a binomial random variable conditional on the hospital'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 five different points: at the minimum number of quality reporting events for the measure (i.e. >10 PCI procedures, which essentially includes all centers); at the mean number of quality reporting events per hospital; and above the 25th, 50th and 75th percentiles of the number of elective procedures.

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)

Level	2011 Signal-to-Noise Ratio	2012 Signal-to-Noise Ratio
All, >10 Procedures	.772	.776
>25 th percentile of volume	.805	.811
>Mean	.847	.845
>75 th percentile of volume	.892	.890
>Average # of PCI procedures	.881	.880

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 similar for 2011 and 2012. At the minimum number of procedures required (>10) the average reliability was .772 and .776, respectively for 2011 and 2012. Among hospitals performing greater than the mean number of elective procedures, the reliability was higher at .847 and .845 in 2011 and 2012, median.

A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of 1.0 implies that all the variability is attributable to actual differences in performance. A reliability of 0.70 is generally considered a minimum threshold for reliability and 0.80 is considered very good reliability.

This measure has moderate reliability across all centers and high reliability above the mean number of elective procedures (50th percentile), where the signal-to-noise ratios was 0.847 and 0.845 in 2011 and 2012, respectively. This suggests that for hospitals with an average or greater number of elective PCI procedures (503 PCI procedures) the measure has high reliability.

2b2. VALIDITY TESTING

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

- ☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☐ **Performance measure score**
 - ☐ **Empirical validity testing**
 - ☒ **Systematic assessment of face validity of performance measure score as an indicator of quality or resource use** (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

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

Content validity of the measure: In a fee-for-service healthcare system, where there have been rapidly escalating costs for the past several decades, there is a pressing need to be able to assess whether or not the risks and costs of a particular procedure justify its use. Against this backdrop, the ACC and numerous other professional societies have banded together to rigorously develop Appropriate Use Criteria (AUC)(1). One set of AUC was developed for coronary revascularization.(2) The ACC NCDR Cath/PCI registry was specifically redesigned in order to be able to capture the indications for PCI and to be able to provide benchmarked reports to hospitals so that they can assess the quality of their medical decision-making. While there are few inappropriate procedures in the setting of acute coronary syndromes, almost 1 in 8 elective procedures are inappropriate,(3) with wide variation across hospitals. This represents an important opportunity to improve medical decision-making and quality. However, in order for such a quality improvement opportunity to be realized, it is critical that each elective PCI procedure have the requisite data to be able to map the patient to an AUC indication. This measure quantifies the proportion of each practice's patients that cannot be mapped to the AUC. A lower percentage rate is better for this measure.

An interesting caveat of this measure is that some may argue, on clinical grounds, that pre-procedure risk stratification with stress testing might not be warranted when the clinical suspicion of coronary disease is so high that the most efficient method for treating patients is to proceed directly to angiography and revascularization. To assess this, NCDR investigators compared the prevalence of significant coronary disease (left main stenosis >50% or any other vessel with a >70% stenosis) among those who could and could not be mapped (where finding a much higher prevalence of significant CAD or higher symptom burden would justify the clinical concern). In these analyses, currently under journal review, we found that among 797,870 patients without a prior history of CAD, the 37.5% without prior stress testing were more likely to be asymptomatic (40.0% vs. 28.2%, $p<0.001$) and less likely to have obstructive coronary disease (35.1% vs. 40.1% $p<0.001$) – the exact opposite of what would be expected if not providing the data to assign an AUC rating were to be associated with better clinical decision-making.(4)

Face validity of the measure: Beyond the clinical logic described above, the ACC, PCPI and AHA systematically assessed the face validity of this proposed measure and a means to improving quality as follows:

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

“The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality.” The respondents recorded their rating on a scale of 1-5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5= Strongly Agree

There were 17 committee members who completed the survey; one respondent was excluded because he was a member of the workgroup that developed this measure. Further information on the survey respondents is available if needed.

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

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

N = 16; Mean rating = 4.5 and 93.8% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 0

3 - 1 (Neither Agree nor Disagree)

4 - 6

5 - 9 (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 very high face validity 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)

There are no exclusions for this measure, in that all elective PCI patients at each hospital are included. While acute coronary syndrome patients are excluded, virtually all of them are mappable to the AUC (<0.3% of acute coronary syndrome PCI patients were not mapped in 2012).

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)

The rates of unmappable acute coronary syndrome patients are shown in the table below:

Non-Elective AUC										
	Total	Quarterly rates from 2011-2012								P-Value
	n = 1021423	2011Q1 n = 122022	2011Q2 n = 125549	2011Q3 n = 122003	2011Q4 n = 124513	2012Q1 n = 133931	2012Q2 n = 132307	2012Q3 n = 130575	2012Q4 n = 130523	
Not Mappable to AUC	2730 (0.3%)	437 (0.4%)	352 (0.3%)	313 (0.3%)	347 (0.3%)	361 (0.3%)	321 (0.2%)	298 (0.2%)	301 (0.2%)	< 0.001

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*)

This measure has been designed to assess the proportion of elective PCI patients in whom the strength of the procedure's appropriateness could be mapped. There are no exclusions for this population and there is no potential bias that could be introduced from this measurement effort.

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 hospital performance on this measure based on sex, age, race and a number of other patient factors, including prior medical history and presenting symptoms to identify variations. Full testing report with information on all patient characteristics tested is available in Appendix A-1.

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

We observed extraordinary variation in this measure across hospitals caring for patients undergoing elective PCI ranging from hospitals in whom none of their elective PCI procedures could be mapped to an AUC indication to other hospitals where all of their elective cases could be mapped. This was observed in both 2011 and 2012. We not only describe the distribution of performance, but also summarize this variation by calculating the median odds ratio (MOR). The MOR comes from a

hierarchical model that adjusts for patient characteristics and examines the variation in the likelihood that one hospital versus another would have comprehensively documented the indications for elective PCI. This can be thought of as the likelihood that a statistically identical patient, presenting to 2 different hospitals in our sample, would have had indications for their PCI documented. These data for 2011 and 2012 are provided below:

Descriptive Statistics in 2011 at Site Level (Sites with 10 or more elective procedures)

Analysis Variable : Proportion Unmappable							
Number of Sites	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Quartile Range
1146	0.4162694	0	0.2800000	0.3953488	0.5403587	1.0000000	0.2603587

By Decile:

10 th Percentile	20 th Percentile	30 th Percentile	40 th Percentile	50 th Percentile	60 th Percentile	70 th Percentile	80 th Percentile	90 th Percentile
0.19008	0.25	0.30827	0.35	0.39535	0.45215	0.50127	0.58416	0.6875

Descriptive Statistics in 2012 at Site Level (Sites with 10 or more elective procedures)

Analysis Variable : Proportion Unmappable							
Number of Sites	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Quartile Range
1178	0.3685528	0	0.2380952	0.3465463	0.4842767	0.9444444	0.2461815

By Decile:

10 th Percentile	20 th Percentile	30 th Percentile	40 th Percentile	50 th Percentile	60 th Percentile	70 th Percentile	80 th Percentile	90 th Percentile
0.15152	0.21739	0.26154	0.30556	0.34655	0.39130	0.45	0.52381	0.62931

A large amount of variability was noted across hospitals. In 2012 the range was 0-100% with the inter-quartile range being 21% to 50%. This yielded a Median Odds Ratio of 2.21 (2.15, 2.26). The Median Odds Ratio (MOR) measures the variation between clusters by comparing the likelihood that 2 statistically identical patients would have undergone the necessary pre-procedural risk stratification to understand the strength of the indication for their procedure were they to present to 2 randomly chosen hospitals. A MOR of 2.21 indicates a large amount of variation among the hospitals, suggesting that the same patient treated at 2 hospitals would have over a two-fold difference in their likelihood of care that achieves this measure.(5) We believe this information indicates a substantial opportunity to improve patient care.

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?)

Given the clinical importance of being able to transparently understand the strength of the indication to perform PCI, the wide variability across hospitals and the high signal-to-noise ratio, we believe that this measure is capable to detecting important differences in the quality of care across hospitals. Moreover, there are very few differences by patient characteristics, as they have little to do with having the documentation available to describe the indications for elective PCI. Thus, the observed variations that we detected are likely due to real differences in performance between sites and represent important opportunities to improve patient care.

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.*

This is not applicable to this measure.

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)*

Missing values are interpreted as 'No' for most variables. However, in general, the CathPCI registry has very little missing data due to its robust data quality program. We interpret missing documentation of indications for elective PCI to indicate a failure to meet the measure and not as 'missing' due to inadequacies of data collection processes. The primary reason for failing to meet the measure ("not mappable" PCIs) is missing stress test information – an essential clinical data element for proper decision-making regarding the appropriateness of PCI.

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 interpretation of missing data to represent a 'failure to meet the measure' (see above), no empirical analysis of the frequency or distribution of missing data was required. For this measure, missing data represented a failure and was included in the primary analysis.

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 high rates of missing data due to a failure to meet the measure. This represents an important opportunity to improve patient care by better documentation prior to PCI of essential clinical data elements required for optimal decision-making.

References

1. Patel MR, Spertus JA, Brindis RG, Hendel RC, Douglas PS, Peterson ED, Wolk MJ, Allen JM, Raskin IE. ACCF proposed method for evaluating the appropriateness of cardiovascular imaging. *J Am Coll Cardiol* 2005;46:1606-13.
2. Patel MR, Dehmer GJ, Hirshfeld JW, Smith PK, Spertus JA. ACCF/SCAI/STS/AATS/AHA/ASNC 2009 Appropriateness Criteria for Coronary Revascularization: a report by the American College of Cardiology Foundation Appropriateness Criteria Task Force, Society for Cardiovascular Angiography and Interventions, Society of Thoracic Surgeons, American Association for Thoracic Surgery, American Heart Association, and the American Society of Nuclear Cardiology Endorsed by the American Society of Echocardiography, the Heart Failure Society of America, and the Society of Cardiovascular Computed Tomography. *J Am Coll Cardiol* 2009;53:530-53.
3. Chan PS, Patel MR, Klein LW, Krone RJ, Dehmer GJ, Kennedy K, Nallamothu BK, Weaver WD, Masoudi FA, Rumsfeld JS, Brindis RG, Spertus JA. Appropriateness of percutaneous coronary intervention. *JAMA*. 2011;306:53-61. PMID: 3293218
4. Abdallah MS, Spertus JA, Mercado N, Nallamothu BK, Kennedy K, Arnold SV, Chan PS. Clinical Symptoms and Angiographic Findings of Patients Undergoing Percutaneous Coronary Intervention without Prior Stress Testing: Insights from the NCDR. *Circ Cardiovasc Qual Outcomes*. 2012; 5: A251
5. Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol*. 2005 Jan 1;161(1):81-8. PubMed PMID: 15615918