

## NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b2-2b6, 2d)

**Composite Measure Title:** PCI: Post-Procedural Optimal Medical Therapy Composite

**Date of Submission:** 12/23/2013

### Composite Construction:

- ☐ Two or more individual performance measure scores combined into one score
- ☒ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)
- ☐ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient) **Instructions: Please contact NQF staff before you begin.**
  - If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
  - 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 composite measures, sections 1, 2a2, 2b2, 2b3, 2b5, and 2d must be completed.**
  - **For composites with 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), validity (2b2-2b6), and composites (2d) 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 25 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** <sup>10</sup> 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.

**2b2. Validity testing** <sup>11</sup> 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.

**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; <sup>12</sup>

### 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).

**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; <sup>14,15</sup> 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 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.**

**Composite 2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:**

- 1) the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
- 2) the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible; and
- 3) the extent of missing data and how the specified handling of missing data minimizes bias (i.e., achieves scores that are an accurate reflection of quality).

## 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 different components in the composite, indicate the component 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).

We propose to use a clinical registry, the National Cardiovascular Data Registry for CathPCI Registry. This is a national quality improvement registry that is currently utilized by >1200 US hospitals performing percutaneous coronary intervention (PCI). Some states and healthcare systems mandate participation. Rigorous quality standards are applied to the data and both quarterly and ad hoc performance reports are generated for participating centers to track and improve their performance.

**1.3. What are the dates of the data used in testing?** Click here to enter date range

All hospital discharges of patients with PCI performed between Jan 2011 and Dec 2012 in the CathPCI registry were used for testing.

**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)**

- ☒ individual clinician  
☐ group/practice  
☐ hospital/facility/agency  
☐ health plan  
☐ other: Click here to describe

**Measure Tested at Level of:**

- ☒ individual clinician  
☐ group/practice  
☐ hospital/facility/agency  
☐ health plan  
☐ 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)

A total of 11,699 operators were included in the analysis. The mean volume of procedures performed by each operator ranged from 1 to 1333, with a mean of 91.5 (Standard Deviation, 101.5) and a median of 59 (interquartile range, 17-132). Approximately half (49.49%) of the physicians in the sample performed PCIs at only one facility; 28.93% performed PCIs at 2 facilities and 21.58% performed PCIs at more than 2 facilities. For reliability testing, we focused on a sample of 4064 operators with at least 50 procedures.

**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)*

A total of 1,070,136 patient discharges performed by 11,699 operators were included in the analysis. For patients with more than one PCI during a single admission that were performed by different operators, we attributed statistics used for this testing to the last operator to perform PCI. This assumption is consistent with the goal of the measure as it is trying to capture appropriate care at the time of discharge. This would be the responsibility of the last operator to perform PCI. See table below for basic demographic characteristics of the patients included in the testing and analysis.

Description	Total	
	#	%
ALL	1070136	100.00
Age>=65		
No	535223	50.01
Yes	534913	49.99
Female		
No	726778	67.91
Yes	343358	32.09
RACE		
Hispanic	57242	5.35
White non-Hispanic	888722	83.05
Black non-Hispanic	86878	8.12
Other	37294	3.48

**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 sample described above was used for general aspects of testing and analysis. For reliability testing, however, we limited the study sample to 4064 operators with at least 50 procedures.

## 2a2. RELIABILITY TESTING

### 2a2.1. What level of reliability testing was conducted?

**Note:** Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score. ☒ **Performance measure score** (e.g., signal-to-noise analysis)

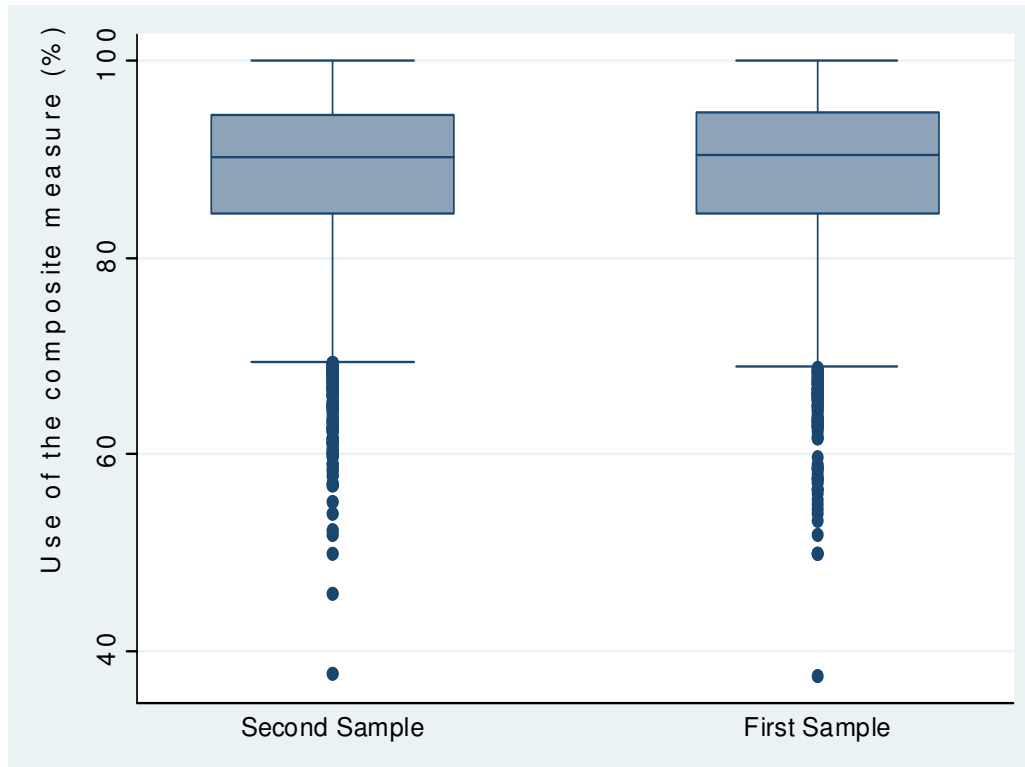
### 2a2.2. Describe the method of reliability testing and what it tests *(describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used)*

To assess reliability, we identified all operators who had performed at least 50 PCIs during the study period. We randomly split patients from this group of operators into 2 samples and then examined the correlation between measures from the 2 samples. Operators who performed less than 50 PCIs during the study period were excluded since the purpose of this analysis was to examine reliability and the small number of procedures would produce less stable estimates of reliability. Standard statistics, such as the mean and standard deviation as well as median and inter-quartile range, were compared. A correlation coefficient also was calculated. The correlation coefficient quantifies how similar estimates were in the 2 groups and provides an assessment of reliability.

**2a2.3. 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)

The correlation coefficient for the two samples was 0.826. See details in table and figure below.

Distribution of The Composite Measure at Discharge Stratified by The Randomly Split Samples		
Description	Randomly Split Samples	
	First (RAND=1)	Second (RAND=0)
	DCM	DCM
N	4044	4064
Mean	0.8867	0.8867
Std Deviation	0.0811	0.0804
75% Q3	0.9481	0.9464
50% Median	0.9033	0.9027
25% Q1	0.8437	0.8453
DCM: Performance rate on the composite measure		
Among physicians with at least 50 cases		
<b>Correlation coefficient: 0.82626</b>		



#### 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?)

Overall, we interpret these results to indicate that the measure is reliable. It distinguishes meaningful differences in performance among operators and is not susceptible to “noise” with random variation. The box and whisker plot of the distribution of operator performance for the composite measure at discharge show a similar percentage of use of the composite measure at discharge for both samples. Furthermore, the correlation coefficient is demonstrated in the figure and shows a strong positive association between both samples.

There are other aspects of reliability regarding data collection that also should be considered. The NCDR Data Quality Program ensures that data submitted to the NCDR are complete validly collected. The NCDR Data Quality Program consists of 3 main components: data completeness, consistency, and accuracy. Completeness focuses on the proportion of missing data within fields, whereas consistency determines the extent to which logically related fields contain values consistent with other fields. Accuracy characterizes the agreement between registry data and the contents of original charts from the hospitals submitting data. Before entering the Enterprise Data Warehouse (EDW), all submissions are scored for file integrity and data completeness, receiving 1 of 3 scores that are transmitted back to facilities using a color coding scheme. A “red light” means that a submission has failed because of file integrity problems such as excessive missing data and internally inconsistent data. Such data are not processed or loaded into the EDW. A “yellow light” status means that a submission has passed the integrity checks but failed in completeness according to predetermined thresholds. Such data are processed and loaded into the EDW but are not included in any registry aggregate computations until corrected. Facilities are notified about data submission problems and provided an opportunity to resubmit data. Finally, a “green light” means that a submission has passed all integrity and quality checks. Such submissions are loaded to the EDW. After passing the DQR, data are loaded into a common EDW that houses data from all registries and included for all registry aggregate computations. In a secondary transaction process, data are loaded into registry-specific, dimensionally modeled data marts. A summary of the Program is noted in the table below.

**Table. Data Quality Program Overview for CathPCI Registry**

<b>Methodology</b>	<ul style="list-style-type: none"> <li>• Nationwide program (i.e., all submitting participants in the United States)</li> <li>• Review of data submitted the previous year</li> <li>• Review of a subset of data elements that can rotate each year</li> <li>• Remote review of data combined with couple of onsite visit</li> <li>• Onsite visits are targeted based on the Data Outlier Program</li> <li>• Random selection of sites and records</li> <li>• Blinded data abstraction from medical charts</li> <li>• Inter-rater Reliability Assessment conducted to validate the audit findings</li> <li>• Adjudication step for participant to refute audit findings</li> </ul>
<b>Scope</b>	<ul style="list-style-type: none"> <li>• Review of hospital's medical records for related episodes of care</li> <li>• Assessment of complete submission (Comparison of two lists : hospital list of cases with specific billing codes versus NCDR submitted records)</li> </ul>
<b>Criteria for selecting sites/records</b>	Remote audit : <ul style="list-style-type: none"> <li>• Sites passing their quarterly Data Quality Report for 2 quarters within audited year</li> <li>• Sites submitting at least the number of records/sites being reviewed</li> </ul> Onsite audit <ul style="list-style-type: none"> <li>• Sites identified with an outlier and not contacted with the data outlier program</li> </ul>
<b>Scoring</b>	NCDR uses a grading system for identifying the amount of agreement or matching between the data captured during the medical record review and data submitted to the NCDR.

## 2b2. VALIDITY TESTING

**Note:** Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.

**2b2.1. What level of validity testing was conducted?** ☒ **Composite 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)

☒ **Systematic assessment of content validity**

☐ **Validity testing for component measures** (check all that apply)

**Note:** applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.

☐ **Endorsed (or submitted) as individual performance measures**

☐ **Critical data elements** (data element validity must address ALL critical data elements)

☐ **Empirical validity testing of the component measure score(s)**

☒ **Systematic assessment of face validity of component measure score(s) 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 checked above, describe the method of validity testing and what it tests** (describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements

*compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Content validity for this measure was systematically assessed by expert work group members during the development process during extensive discussion and a final confidential vote. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and concurrent formal peer review process. Additionally, comments were solicited from a panel of consumer, purchaser, and patient representatives convened by the AMA-PCPI specifically for this purpose. All comments received were reviewed by the expert work group and the measures were adjusted as needed. Finally, the measure underwent review and approval by the Board of Trustees of the ACC and the Science Advisory and Coordinating Committee of the AHA, as well as review and voting by the PCPI membership.

Members of the expert work group that developed the measure included: Brahmajee K. Nallamothu, MD, MPH, FACC, FAHA, *Co-chair*; Carl L. Tommaso, MD, FACC, FAHA, FSCAI, *Co-chair*; H. Vernon Anderson, MD, FACC, FAHA, FSCAI; Jeffrey L. Anderson, MD, FACC, FAHA, MACP; Joseph C. Cleveland, Jr., MD; R. Adams Dudley, MD, MBA; Peter Louis Duffy, MD, MMM, FACC, FSCAI; David P. Faxon, MD, FACC, FAHA; Hitinder S. Gurm, MD, FACC; Lawrence A. Hamilton; Neil C. Jensen, MHA, MBA; Richard A. Josephson, MD, MS, FACC, FAHA, FAACVPR; David J. Malenka, MD, FACC, FAHA; Calin V. Maniu, MD, FACC, FAHA, FSCAI; Kevin W. McCabe, MD; James D. Mortimer; Manesh R. Patel, MD, FACC; Stephen D. Persell, MD, MPH; John S. Rumsfeld, MD, PhD, FACC, FAHA; Kendrick A. Shunk, MD, PhD, FACC, FAHA, FSCAI; Sidney C. Smith, Jr., MD, FACC, FAHA, FACP; Stephen J. Stanko, MBA, BA, AA and Brook Watts, MD, MS.

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

After the measure was fully specified, members of two existing committees, one at the ACC, one at AHA and one joint ACC/AHA, with expertise in 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 = 4.44 and 87.5% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

### **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 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. The individual components have been associated with better outcomes and are accepted quality measures in patient populations. As noted in Section 2a2.4., we also have good evidence of validity from the CathPCI audit data. A vast majority of data elements showing >90% agreement between routinely collected data and audit data. (See CathPCI bleeding and mortality applications).



## 2b3. EXCLUSIONS ANALYSIS

**Note:** Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.

**NA** ☐ no exclusions — skip to section 2b4

Exclusions in our measure are appropriate and intended to remove patients who would not be expected to be prescribed all medications for which they are eligible at discharge. Specifically, we exclude patients who expired; patients who left against medical advice; patients discharged to hospice or for whom comfort care measures only is documented and patients discharged to other acute care hospital. Because this is a composite, some patients are eligible for some of the medications but not for others. This variation is incorporated into the construction of the composite measure. The overall frequency and proportion of the discharges excluded for each exclusion criterion is displayed below.

**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)

The majority of exclusions for this measure are noted below (discharge status of expired; discharge location of “other acute hospital, hospice, or against medical advice”). These exclusions are relatively rare and firmly supported by the clinical rationale.

**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)

Although not an exclusion for this measure, we needed to remove patients from the testing samples for this data source due to a lack of sufficient information on operator identification. This made up approximately 14% of the study population. This is not an exclusion criterion that will be utilized when the performance measure is put in practice but was only done to ensure that the statistics performed to assess testing were robust.

<b>Discharge Status: deceased</b>	<b>19617</b>	<b>1.53</b>
Remaining	<b>1262743</b>	<b>98.47</b>
<b>Discharge Location: Other acute care hospital</b>	10937	0.87
Remaining	<b>1251806</b>	<b>99.13</b>
<b>Discharge Location: Hospice</b>	2209	0.18
Remaining	<b>1249597</b>	<b>99.82</b>
<b>Discharge Location: Left against medical advice</b>	3078	0.25
Remaining	<b>1246519</b>	<b>99.75</b>
<b>Not eligible to the composite measure</b>	180	0.01
Remaining	<b>1246339</b>	<b>99.99</b>
<b>NPI unknown</b>	171178	13.73
Remaining	<b>1075161</b>	<b>86.27</b>
<b>NPI invalid*</b>	5025	0.47
Study Sample	<b>1070136</b>	<b>99.53</b>

**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 all exclusions is very low with the exception of exclusions because of a lack of operator identification numbers (i.e. National Provider Identification [NPI]). While this may be a challenge within this data source, we do not believe it will be a challenge for real-world implementation of this performance measure since operator identification numbers may be readily identified using available information for most health systems.

Thus, we argue that there are no 'discretionary' exclusions. All exclusions are necessary for the accurate calculation of performance on the composite measure. For example, patients need to survive to discharge to be eligible for the measure. Similarly, it would be inappropriate to calculate the measure among patients discharged to another acute care facility or those who left the hospital against medical advice. In light of the lack of randomized trials designed to evaluate the efficacy of clopidogrel (P2Y12 receptor blockers) in addition to aspirin compared to aspirin alone in STEMI patients treated with primary PCI, we feel no additional patients should be excluded from our composite measure. The value of including these patients and the potential for evaluating their outcomes in our bleeding and mortality measures outweighs the burden of increased data collection and analysis.

Indirect evidence of long-term benefit exists from trials PCI-CURE, CREDO, and CURE (1, 2) of patients with non-STEMI in which P2Y12 receptor blockers were continued for 9 to 12 months. At 30 days after PCI, clopidogrel therapy was associated with a significant reduction in the primary endpoint of cardiovascular death, MI, or stroke (3.6 versus 6.2 percent, adjusted odds ratio 0.54, 95% CI 0.35-0.85).

Accordingly, we do not believe that additional testing is necessary.

## **2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**

**Note:** Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement. **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? (check all that apply)**

☐ **Endorsed (or submitted) as individual performance measures**

☒ **No risk adjustment or stratification**

☐ **Statistical risk model**

☐ **Stratification by risk categories**

☐ **Other, Click here to enter description**

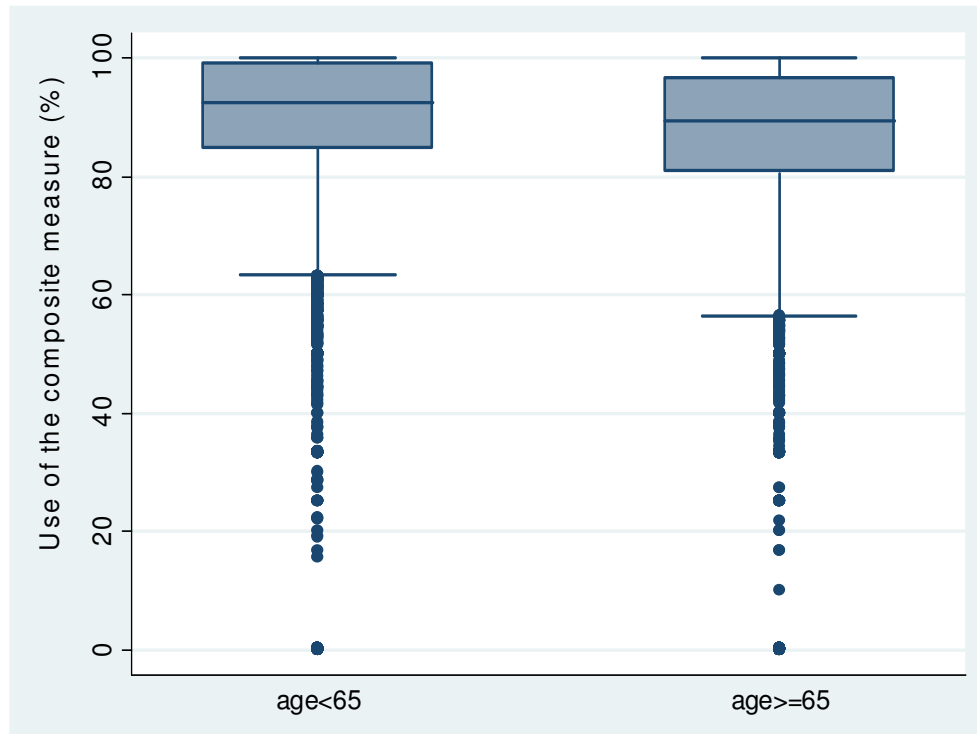
### **2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.**

We evaluated the performance measure across age groups (<65, >65 years), gender, and racial/ethnic categories. We found significant overlap in the results of the performance measure for these different groups. For this reason, we do not recommend that these be separately reported. The figures below indicate these different categories and the distribution of the performance measure among the groups.

#### **Distribution of The Composite Measure at Discharge Stratified by Age**

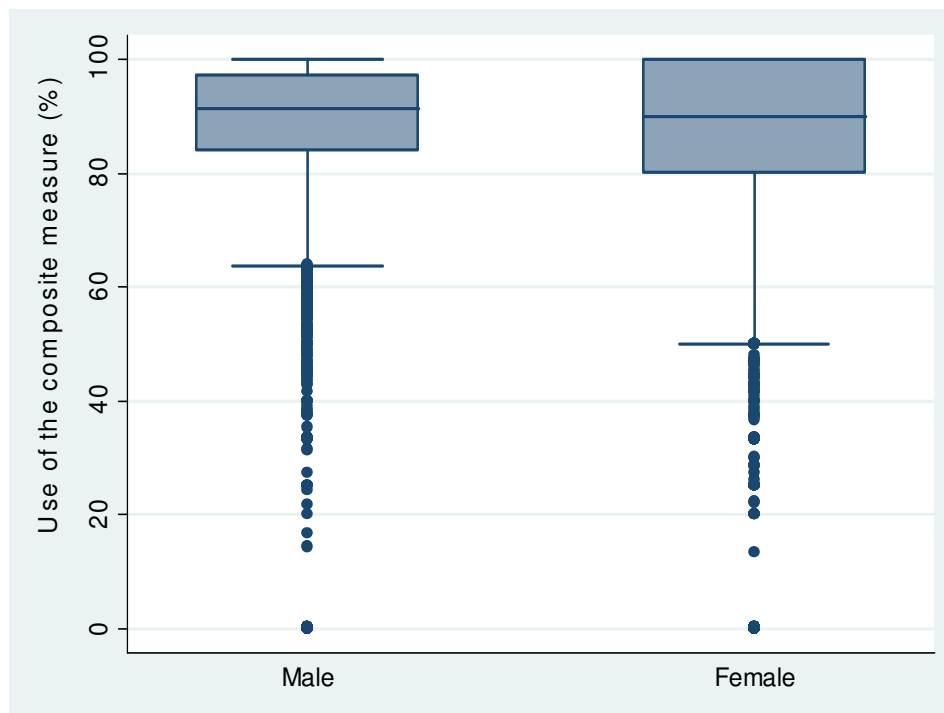
Description	Age >= 65	
	Yes	No

<b>N</b>	11308	11318
<b>Mean</b>	0.8635	0.8912
<b>Std Deviation</b>	0.1462	0.1344
<b>75% Q3</b>	0.9677	0.9922
<b>50% Median</b>	0.8947	0.9249
<b>25% Q1</b>	0.8059	0.8485



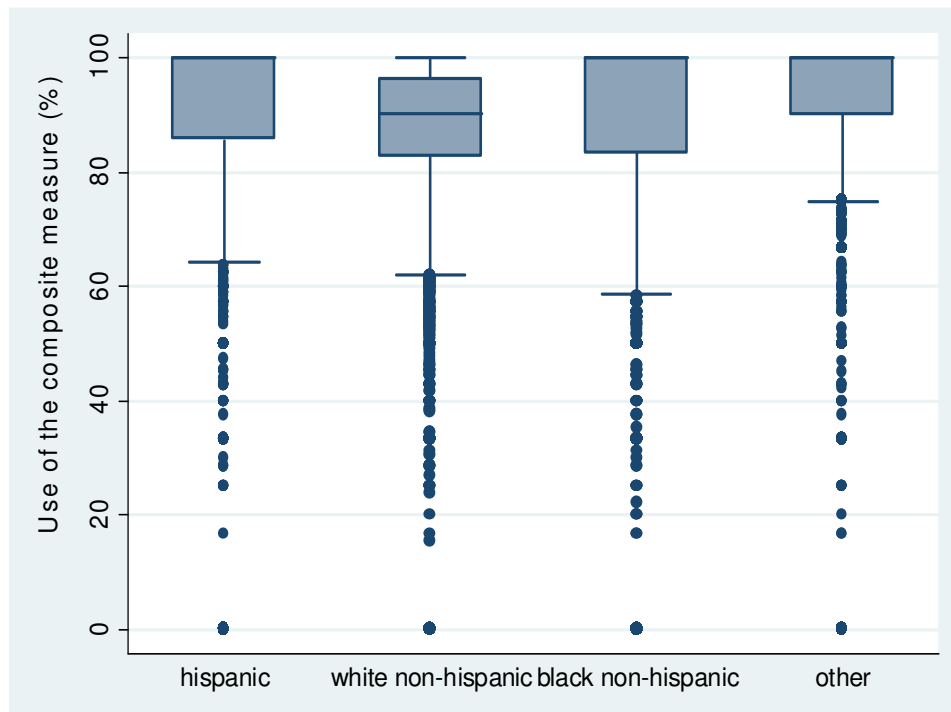
#### Distribution of The Composite Measure at Discharge Stratified by Gender

Description	Female	
	Yes	No
<b>N</b>	11037	11509
<b>Mean</b>	0.8640	0.8831
<b>Std Deviation</b>	0.1567	0.1322
<b>75% Q3</b>	1.0000	0.9725
<b>50% Median</b>	0.9000	0.9143
<b>25% Q1</b>	0.8000	0.8387



Distribution of The Composite Measure at Discharge Stratified by Race

Description	Race			
	Hispanic	White non-Hispanic	Black non-Hispanic	Other
<b>N</b>	7161	11514	8112	6516
<b>Mean</b>	0.8888	0.8760	0.8777	0.8940
<b>Std Deviation</b>	0.2196	0.1333	0.2082	0.2311
<b>75% Q3</b>	1.0000	0.9655	1.0000	1.0000
<b>50% Median</b>	1.0000	0.9048	1.0000	1.0000
<b>25% Q1</b>	0.8571	0.8276	0.8333	0.9000



**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)

NA

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

NA

**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**

NA

**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):

NA.

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

NA.

**2b4.9. Results of Risk Stratification Analysis:**

NA.

**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?)

NA.

**\*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)

NA.

---

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

**Note:** Applies to the composite performance measure.

**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 operator performance for the composite measure to identify meaningful differences. We believe these differences are clinically meaningful as operators in the lowest quartile of performance showed substantial and meaningful differences in their performance on this measure when compared with operators in the top quartile.

**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 also examined differences across categories of age, gender and race/ethnicity. Additional information on the analyses performed is available upon request.

**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?)

The wide gap in performance rates across operators demonstrates that this measure is necessary to improve the quality gap. The lack of large differences across specific age, gender and racial/ethnic groups suggests that stratification is likely not needed.

---

## **2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

**Note:** Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.

***If only one set of specifications for each component, 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*)

NA

**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*)

NA

**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?*)

NA

---

## **2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH**

**Note:** *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

**2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.**

We believe the content validity of this measure has been achieved by virtue of the noted expertise of those individuals who developed this measure. The individual components of the composite have already shown to impact clinical outcomes. However the empirical analysis demonstrating the individual component measures fit the overall quality construct is currently being researched. The testing will focus on construct validation which will test the hypothesis on the theory of the construct that following these processes for patients undergoing PCI lead to better outcomes. This research is expected to ultimately be published in the medical literature. While the analysis will likely not be ready prior to the submission deadline of the Cardiovascular Endorsement Maintenance project, they will be available prior to the close of the measure cycle. The analysis in preparation for publication can be provided upon request or at publication.

**2d1.1 Describe the method used** (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification*)

NA.

**2d1.2. What were the statistical results obtained from the analysis of the components?**(*e.g., correlations, contribution of each component to the composite score, etc.; if no empirical analysis, identify the components that were considered and the pros and cons of each*)

NA.

**2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite?***(i.e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

**2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible**

**NA.**

**2d2.1 Describe the method used***(describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

**NA.**

**2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules?***(e.g., results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each)*

**NA.**

**2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct?***(i.e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting)*

**NA.**

**2d3. Empirical analysis demonstrating that the approach for handling missing data minimizes bias***(i.e., achieves scores that are an accurate reflection of quality).*

**Note:** *Applies to the overall composite measure; the focus is on missing data rather than exclusions, which are considered in 2b3.*

**2d3.1. What is the overall frequency of missing data and the distribution of missing data across providers?**

Discharges with missing or invalid operator identification numbers (NPIs) were excluded from this analysis. This is primarily a current limitation of the registry. Although rates of valid NPI collected are improving (now over 85%), this data was previously not collected universally and permitted entry of invalid numbers; therefore, a portion of the discharges in the registry do not include this data.

**2d3.2. Describe the method used to compare approaches for handling missing data***(describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

**NA.**

**2d3.3. What were the statistical results obtained from the analysis of missing data?** *(e.g., results of sensitivity analysis of effect of various rules for missing data; if no empirical analysis, identify the approaches for handling missing data that were considered and pros and cons of each)*

**NA.**



**2d3.4. What is your interpretation of the results in terms of demonstrating that the approach used for missing data minimizes bias?***(i.e., what do the results mean in terms of supporting the selected approach for missing data; if no empirical analysis, provide rationale for the selected approach for missing data)*

**NA.**

### **References**

1. Mehta SR, Yusuf S, Peters RJ, et al. Effects of pretreatment with clopidogrel and aspirin followed by long-term therapy in patients undergoing percutaneous coronary intervention: the PCI-CURE study. *Lancet*. 2001;358:527-33.
- 2, Steinhubl SR, Berger PB, Brennan DM, et al. Optimal timing for the initiation of pre-treatment with 300 mg clopidogrel before percutaneous coronary intervention. *J Am Coll Cardiol*. 2006;47:939-43.
3. Yusuf S, Zhao F, Mehta SR, et al. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *N Engl J Med*. 2001;345:494-502.