**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0643

**Measure Title**: Cardiac Rehabilitation Patient Referral from an Outpatient Setting

**Date of Submission**: 11/8/2017

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | 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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 25 pages (*incuding 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](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing 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**](#Note10) 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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) 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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **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**](#Note13)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** 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**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

The following datasets were used: AACVPR/ACCF/AHA Cardiac Rehabilitation Referral and Reliability (CR3) Project and the ACCF/AHA PINNACLE Registry

**1.3. What are the dates of the data used in testing**? 2009-2012

**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.20*)** | **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*)

**AACVPR/ACCF/AHA CR3 Project:**

U.S. practices identified from the ACCF, AHA, and AACVPR databases were invited to participate in the Cardiac Rehabilitation Referral and Reliability (CR3) Project. We sought a variety of outpatient practices, based on varied geographical locations, community sizes, and hospital types/sizes. Outpatient practices that met participation criteria were included in the project. Participation criteria included a willingness and ability to: (1) provide a study coordinator and 2 chart abstractors, (2) complete the project within the specified timeline, and (3) obtain local IRB approval to carry out the project in their outpatient practice. Once each hospital completed and submitted their required data, they were sent a small token of appreciation from AACVPR, ACCF, and AHA. A total of 45 outpatient practices expressed an interest in participating in the project, including hospitals from outside the U.S. (Puerto Rico, Romania, and Turkey). 6 outpatient centers (all in the United states and distributed around the country) met all participation criteria and were selected to participate in the project. The sites used a mixture of paper medical records and EHR systems.

**ACCF PINNACLE Registry :**

Data were analyzed from the ACCF outpatient registry, PINNACLE. The sample populations, for calendar year 2011 and calendar year 2012, include 252,331 patients from 994 practice in 2011 and 298,206 patients from 1022 practices in 2012.

**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*)

**AACVPR/ACCF/AHA CR3 Project:**

Descriptive statistics are noted below:

**Sex:** Male: 65% (n=152 / 234), Female: 35% (n=82/234)

**Age:** 18-39: 3% (n=7/229), 40-64: 40% (n=91 / 229 ), 65-79: 45% (n=103 / 229) , 80+: 12% (n=28 / 229 )

**Race:** White: 84% (n=196 / 234), Black: 8% (n=19 / 234), Asian: 0% (n=1 / 234), American Indian: 1% (n=3 / 234), Native Hawaiian/Pacific Islander: 0% (n=1 / 234), Other: 6% (n=14 / 234)

**Hispanic Ethnicity:** 0% (n=1 / 234)

**PINNACLE Registry, 2012:**

|  | **Total** |
| --- | --- |
| **n = 252331** |
| Race   (1) White   (2) Black   (3) Other   Missing (.) | 117261 (  89.9% ) 8758 (   6.7% ) 4415 (   3.4% ) 121897 |
| Insurance   (0) No insurance   (1) Private   (2) Medicare   (3) Medicaid   (4) Other   Missing (.) | 14914 (   7.0% ) 129907 (  61.1% ) 61289 (  28.8% ) 3956 (   1.9% ) 2629 (   1.2% ) 39636 |
| Age   18 to <60   60 to <70   70 to <80   80 to 112 | 71020 (  28.1% ) 67696 (  26.8% ) 65497 (  26.0% ) 48118 (  19.1% ) |
| Sex   (1) Male   (2) Female   Missing (.) | 149415 (  59.2% ) 102812 (  40.8% ) 104 |
| BMI   Missing | 29.7 ± 6.4 91870 |
| Diabetes | 66294 (  26.3% ) |
| CAD | 247440 (  98.1% ) |
| Hypertension | 209013 (  82.8% ) |
| AFib | 59525 (  23.6% ) |
| HF | 76388 (  30.3% ) |
| PAD | 89780 (  35.6% ) |
| Prior Stroke/TIA | 79532 (  31.5% ) |
| MI history | 125549 (  49.8% ) |

**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 datasets described above was used for all aspects of testing.

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*)

**AACVPR/ACCF/AHA CR3 Project:**

The aim of this project was to assess the reliability and feasibility of abstracting the Cardiac Rehabilitation Referral Performance Measure from an outpatient setting. The sites identified to participate in the project were asked to identify one study coordinator and two chart abstractors. 35 patients were identified by the study coordinator at each site from a consecutive sample of patients admitted to their hospital having a qualifying diagnosis for CR, and discharged alive, starting in August 1, 2009. The two abstractors at each site reviewed the same 35 patient records from their hospital twice (once at baseline, and again one week later).

Site coordinators were instructed to include in the 35 patient records, 30 patients who had qualifying diagnoses for CR/SP referral (to capture sensitivity testing) and 5 patients who did not have a qualifying diagnosis for CR/SP referral (to capture specificity testing). The qualifying events are indicated in the measure numerator statement. The non-qualifying events for the purpose of this abstraction project needed to have one or more of the following diagnoses: heart failure, atrial fibrillation, or syncope.

The CR3 Project Workgroup worked with the study coordinators to address reliability, feasibility, and usability properties for the cardiac rehabilitation performance measures. Specifically the workgroup created chart abstraction forms, site coordinator instructions, abstractor instructions, sample IRB protocol, frequently asked questions, and tracking forms to keep track of the intra-rater (1 abstractor reviewing the same patient record two times) and inter-rater process (2 abstractors reviewing the same patient record). The workgroup had a kickoff call with each center’s study coordinator to acquaint him/her with the abstraction project. The workgroup communicated weekly with site coordinators to address any questions or comments the sites may have had.

Abstractors reviewed each patient record and completed the CR3 Project form (see supplement).

Definitions used:

Eligible patient: a patient that had a qualifying event/diagnosis during the hospital period under review

Non-eligible patient: a patient that did not have a qualifying event/diagnosis during the hospital period under review

CR/SP referral : documentation in the patient record for the index hospitalization that the patient was being referred to an outpatient cardiac rehabilitation/secondary prevention program

Exception to referral : documentation in the patient record for the index hospitalization that a patient who was eligible for CR/SP referral had a patient, medical, or healthcare system exception that prohibited their participation in CR/SP

Analyses were performed as follows:

1. Intra-rater and inter-rater agreement between patient record reviews

Eligibility: Was the patient eligible for CR/SP referral?

CR/SP Referral: Was each eligible patient referred to CR/SP?

Exceptions: For patients not referred to CR/SP, was/were any exception(s) to CR/SP documented?

1. Percent agreement

In what percentage of patient record abstractions did the abstractors agree (for both intra-rater and inter-rater agreement)?

1. Kappa statistic

Site specific: Calculated for the 2 abstractors at each site, to compare intra- and inter-abstractor reliability, with regards to his/her assessment of: (1) eligibility for CR/SP referral, (2) referral to CR/SP, and (3) exceptions to CR/SP referral

Pooled estimate: data from all sites were combined to calculate a pooled kappa statistic for intra- and inter-observer reliability for assessing CR/SP eligibility, referral, and exceptions.

By convention, a kappa > .70 is considered acceptable inter-rater reliability.(1) We used the scale below for our analysis.

0: No better than chance

0.01-0.20: Slight

0.21-0.40: Fair

0.41-0.60: Moderate

0.61-0.80: Substantial

0.81-1.0: Almost perfect

(Reference: Landis J, Koch G, The measurement of observer agreement for categorical data, *Biometrics*, 1977;33:159-174.)

It is important to consider both the “percent agreement” and the kappa statistic when assessing the reliability of abstracting this performance measure from patient records, especially for the assessment of “eligibility” and “exceptions”. Each method of reliability assessment gives a slightly different view of reliability in this case.

“Percent agreement” is a helpful assessment of reliability of the measure, but given that over 80% of the patients in the study sample were eligible for cardiac rehabilitation, and more than 90% of the patients were free from exceptions to cardiac rehabilitation participation, the percent agreement for the abstractors may have been somewhat inflated, since by chance alone abstractors may have chosen the “right” eligibility or exception status. (To help minimize this, we blinded the abstractors to the actual number/percentage of patients who were eligible for cardiac rehabilitation in their sample. In addition, abstractors were unaware of the range of exceptions that would be expected in their sample.)

The kappa statistic performs best when there is nearly equal chance of study outcomes (for example, equal chance of being eligible or not eligible for cardiac rehabilitation). When there is a high likelihood of one of the two outcomes, as in our study (high likelihood of eligibility), the results of the kappa analyses can sometimes be less accurate and actually underestimate the true reliability the measure due to a phenomenon that is referred to as a “kappa score paradox” in which there is high percent agreement, yet a low kappa score. (Reference: Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes*.* [*J Clin Epidemiol.*](http://www.ncbi.nlm.nih.gov/pubmed/8621993##) 1996 Apr;49(4):431-4.) Indeed, we observed in our site specific analyses that in some centers with very high percent agreement within and between abstractors, the kappa statistics were very low or even zero in some rare cases. With this in mind, the kappa statistic may underestimate the true reliability of the CR measure.

Using both the “percent agreement” and the kappa statistic together provides a robust view of the reliability of the CR performance measure. One (“percent agreement”) may slightly overestimate reliability and the other (kappa statistic) may slightly underestimate reliability. The true reliability of the measure most likely lies between the results from the two methods of assessment. Since the “percent agreement” method suggests “almost perfect” reliability and the kappa statistic suggests “substantial” to “almost perfect” reliability, the overall reliability of the CR performance measure appears to be between “substantial” and “almost perfect”

**PINNACLE Registry:**

Data were used to assess reliability and other performance characteristics for centers participating in the PINNACLE Registry from January 1 2011 until December 31, 2012.

Reliability of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician ) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician’s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

Reliability is estimated five different points: at the minimum number of quality reporting events for the measure; at the mean number of quality reporting events per physician; and at the 25th, 50th and 75th percentiles of the number of quality reporting events.

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

**AACVPR/ACCF/AHA CR3 Project (pooled data results):**

The abstractor and coordinator experiences in chart abstraction prior to participating in the CR3 project varied greatly. The summary data is below.

Less than 1 month 39% (n= 11)

1-6 months 11% (n=3)

6-12 months 7% (n=2)

1-2 yrs 4% (n=1)

2-3 yrs 4% (n=1)

3-4 yrs 11% (n=3)

4-5 yrs\_\_\_None

More than 5 years 25% (n=7)

ARE PATIENTS ELIGIBLE FOR CARDIAC REHABILITATION?

Percentage deemed eligible for cardiac rehabilitation: 199 / 234 (85%) (mean of all observations)

(Actual percentage of patients who were eligible for cardiac rehabilitation: 200/234 (86%))

Intra-rater reliability (agreement within the same abstractor):

% Agreement: 232 / 232 (100%)

Kappa: 1.00 (-)

Inter-rater reliability (agreement between abstractors):

% Agreement: 218 / 231 (94%)

Kappa: 0.77 (0.65, 0.89)

HAVE PATIENTS BEEN REFERRED TO CARDIAC REHABILITATION?

Percentage referred to cardiac rehabilitation: 111 /185 (60%) (mean of all observations)

Intra-rater reliability:

% Agreement: 172 / 176 (98%)

Kappa: 0.95 (0.90, 0.99)

Inter-rater reliability:

% Agreement: 148 / 172 (86%)

Kappa: 0.70 (0.59, 0.81)

ARE THERE EXCEPTIONS NOTED FOR ELIGIBLE PATIENTS NOT REFERRED TO CARDIAC REHABILITATION?

Percentage with documented exceptions to cardiac rehabilitation: 17 /201 (9%) (mean of all observations)

Intra-rater reliability:

% Agreement: 189 / 196 (96%)

Kappa: 0.76 (0.60, 0.93)

Inter-rater reliability:

% Agreement: 185 / 191 (97%)

Kappa: 0.79 (0.63, 0.95)

**PINNACLE, 2012:**

| **Description** | **Number of Patients** | **Signal-to-Noise Ratio** |
| --- | --- | --- |
| Minimum | 10 | 0.990 |
| 25th percentile | 87 | 0.995 |
| 50th percentile | 173 | 0.998 |
| 75th percentile | 379 | 0.998 |
| Average | 292 | 0.998 |

**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 CR3 project demonstrates high to very high reliability of the measure. The PINNACLE data analysis demonstrates excellent reliability when evaluated at the minimum level of quality reporting events and higher reliability at the median number of events (50th%), and at average and greater number of quality events.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b1. VALIDITY TESTING**

**2b1.1. 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- Will aim to obtain additional empirical validity testing data for future measures as time allows** **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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. 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)*

**Face and Content Validity**

**Validity Survey of Experts:** Validity of the measure score was systematically assessed as follows: After the measure was fully specified, members of 3 existing committees, one at the ACC, one at AHA and one joint ACC/AHA, with expertise in general cardiology, cardiac rehabilitation, quality improvement, outcomes research, 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

Face and content validity of the measure score was systematically assessed according to responses received from survey respondents.

**AACVPR/ACCF/AHA performance measures set:** To determine the content/context validity of the measures, a process using a Delphi peer review was utilized. An explicit and standardized process for ACCF/AHA performance measure development was followed, including the following steps: 1. Formation of the Development Committee, 2. Identification of Potential Factors, for Inclusion, 3. Scoring of the Factors/Expert Opinion, 4. Public Comment Period/Peer Review, 5. Further Refinement, 6. Final approval by organizations, 7. Peer Review Publication/Endorsement. Reviewers were asked to provide comments on the document on the basis of the rating form and guide shown on page 1432 at Http://content.onlinejacc.org/cgi/reprint/j.jacc.2007.04.033v1.pdf

Content/context validity of the measures was also established by virtue of the specialized expertise of the Performance Measures Work Group members including the structured discussions that the work group conducted, as well as the rigorous peer review and public comment period that were carried out. For this particular topic those individuals who were involved in identifying and drafting the performance measures were leaders and experts in the field of cardiac rehabilitation as chosen by AACVPR, ACCF, and AHA.

Furthermore, additional face and content validity was demonstrated from the update of the measure in 2010. During the NQF Care Coordination project, the Steering Committee asked AACVPR, ACCF, and AHA to remove patient refusal as an exception. Since that time, all 3 organizations have published an updated document (NQF measures 0642 and 0643) that explicitly notes that patient refusal should not be an allowable exception. In addition, the cardiac rehabilitation referral measures were revised to facilitate the implementation of these two measures by including administrative codes to identify denominator-eligible populations. All changes were approved by the American Association of Cardiovascular and Pulmonary Rehabilitation Board of Directors, the American College of Cardiology Foundation Board of Trustees, and by the American Heart Association Science Advisory and Coordinating Committee. The performance measure set was also reviewed via AHA and ACC processes as well as by the AACVPR Document Oversight Committee.

**AACVPR/ACCF/AHA CR3 Project:** Through the NQF endorsement process, the cardiac rehabilitation referral performance measures (“Set A” measures) received time-endorsed status in 2010, thus supporting the content validity of these measures.

**PINNACLE Registry:** ACCF and AHA registries always attempt to include ACCF/AHA Task Force on Performance Measures in their various modules. The measures have content/context validity based on the approach articulated under the AACVPR/ACCF/AHA performance measure set.

**Predictive Validity**

Published data have shown that as the cardiac rehabilitation referral measure is met (i.e., the patient is referred to cardiac rehabilitation), the proximate desired outcome (cardiac rehabilitation participation) increases, as does the longer term desired outcome (reduction in morbidity and mortality rates). For more details, see the supplemental materials.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
There were 27 individuals who completed the survey.  Further information on the survey respondents is available if needed. Results of the survey were as follows:

-Average score: 4.83

**-**93% of respondents either agree or strongly agree that the outpatient measure can accurately distinguish good and poor quality.

**AACVPR/ACCF/AHA performance measures set:**

In May 2007 the final peer reviewed publication of the performance measures document was approved by the American Association of Cardiovascular and Pulmonary Rehabilitation Board of Directors, the American College of Cardiology Foundation Board of Trustees and by the American Heart Association Science Advisory and Coordinating Committee. Additionally, the publication was endorsed by the American College of Chest Physicians, American College of Sports Medicine, American Physical Therapy Association, Canadian Association of Cardiac Rehabilitation, European Association for Cardiovascular Prevention and Rehabilitation, Inter-American Heart Foundation, National Association of Clinical Nurse Specialists, Preventive Cardiovascular Nurses Association, and the Society of Thoracic Surgeons. The final document was published *the Journal of the American College of Cardiology* (the official journal of the American College of Cardiology), *the Journal of Cardiopulmonary Rehabilitation and* *Prevention* (the official journal of the American Association of Cardiovascular and Pulmonary Rehabilitation) and *Circulation* (the official journal of the American Heart Association) in September 2007. The document can be found at <http://content.onlinejacc.org/cgi/reprint/j.jacc.2007.04.033v1.pdf>.

**AACVPR/ACCF/AHA CR3Project:**

The cardiac rehabilitation referral measures (NQF measures 0642 and 0643) were revised in 2010 to clarify numerator and denominator exclusion criteria and to facilitate the implementation of these two measures by including administrative codes to identify denominator-eligible populations.

**PINNACLE Registry:**

A review of the measure based on the attributes, of reliability, ease of implementation, appropriate numerator, denominator, and exception specifications was performed. Given that it fulfilled these attributes, the measure was included in the registry. Data from this registry can be seen throughout the submission form and supplemental materials.

**2b1.4. 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?*)

As noted above, our interpretation is that face and content validity has been established for this measure.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b2.1. 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*)

**AACVPR/ACCF/AHA CR3Project:** Reliability of abstracting measure exclusions was tested in the CR3 project (see reliability testing section above). Exclusions or exceptions include patient, medical, and system-based conditions that would preclude the reasonable participation of a patient in a cardiac rehabilitation program (death, residing in an extended care nursing facility, lack of a cardiac rehabilitation program close to where the patient lives, etc.).

**PINNACLE Registry:** Exclusion rates and reasons were assessed from the PINNACLE Registry.

**2b2.2. 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 reliability of abstracting exclusions was high to very high, as shown in the reliability section of this document.

In the PINNACLE Registry, 95% (n=944) of the providers did not report exclusions/exceptions. Among those providers who did report exclusions/exceptions, the mean rate was 29%. Among the patients with exclusions/exceptions, 7.4% were for medical reasons, 63.6 for patient reasons, and 29% for system reasons.

**2b2.3. 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*)

Documentation and assessment of exclusions are very important for this measure, to help reduce the possibility of bias in reporting (i.e., excluding patients who are actually eligible for cardiac rehabilitation referral, in order to improve performance scores). Based on the results of our CR3 project, the time and effort to assess exclusions does not appear to add significant burden (see supplemental materials section for more details on time required to complete abstraction of data for this measure).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. 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***](#section2b5)***.***

**2b3.1. 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

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

**2b3.2. 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**.

Referral to cardiac rehabilitation is appropriate and evidence-based for all patients who have had a qualifying event/diagnosis/procedure. Referral or non-referral is not based on a patient’s level of risk, but rather cardiac rehabilitation is appropriate and evidence-based for all eligible patients no matter what their risk level

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

**2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:**

**Published literature**

**Internal data analysis**

**Other (please describe)**

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

**2b3.5. 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*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b3.9. Results of Risk Stratification Analysis**:

**2b3.10. 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*)

**2b3.11.** **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*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b4.1. 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)*

**2b4.2. 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*)

**2b4.3. 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?*)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

Performance rates were assessed by chart abstraction in the CR3 project. Reliability of that assessment was also performed, as noted in the reliability section above.

In the PINNACLE Registry, performance rates were assessed by decile, to allow for assessment of differences between “low” and “high” performing centers.

**2b5.2. 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*)

See reliability section for results of CR3 Project.

The PINNACLE Registry results are shown below:

| **# of providers** | **# of patients** | **Minimum** | **Lower Quartile** | **Mean** | **Upper Quartile** | **Maximum** | **Quartile Range** | **Std Dev** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1022 | 298206 | 0.00% | 0.84% | 9.18% | 13.0% | 100% | 12.1% | 12.3% |

|  | **Mean** |
| --- | --- |
| Decile 2 | 0.0% |
| Decile 3 | 0.9% |
| Decile 4 | 2.4% |
| Decile 5 | 4.1% |
| Decile 6 | 6.2% |
| Decile 7 | 8.9% |
| Decile 8 | 13.0% |
| Decile 9 | 19.0% |
| Decile 10 | 36.9% |

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

There is wide variation in performance for this measure, documented in the datasets we used. Use of this measure allows for identification of that variation in delivery of cardiac rehabilitation referral. This is important because it provides data from which centers can identify improve upon gaps in care that are identified.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

In the CR3 Project, lack of documentation of a cardiac rehabilitation referral was assumed to represent “no referral made”. In the PINNACLE database, missing values are interpreted as “no” responses. While it is challenging to ascertain a response that is truly “missing” versus one that is truly “No”, we assume that data were missing if all records from a given practice are missing.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Given our assumptions, noted above, we did not conduct an empirical analysis of the frequency or distribution of missing data in the CR3 project.

In the PINNACLE dataset, 1.2% (13/1022) of centers were identified as having missing data and were excluded from analysis from the PINNACLE Registry.

**2b6.3. 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, based on the data listed above, is that the missing data rate is extremely low for our primary measure.