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

**Measure Number** (*if previously endorsed*)**:** 2152

**Measure Title**: Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling

**Date of Submission**: 8/1/2018

**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 data source is Electronic Health Records and Registry data

The data source is 2016 Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

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

The measurement period (data collected from patients seen) was 8/1/2011 through 7/31/2012.

The data are for the time period January 2016 through December 2016 and cover the entire United States.

**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*)

This study captured performance on 97 events, the data were not captured at the physician level, restricting reporting of variation in performance to the organization level only.

We received data from 9,511 physicians reporting on this measure through the registry option for CMS’s PQRS in 2016. Of those, 8,458 physicians had all the required data elements and met the minimum number of quality reporting events (10) for our analysis for a total of 1,660,749 quality events. For this measure, 89 percent of physicians are included in the analysis, and the average number of quality reporting events are 196 for the remaining 1,660,749 events. The range of quality reporting events for 8,458 physicians included is from 10 to 5,579. The average number of quality reporting events for the remaining 11 percent of physicians that aren’t included is 4.

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

There were 1,660,749 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure after exceptions were removed.

**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 same data samples were used for reliability testing and exceptions analysis.

**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.

Patient-level socio-demographic (SDS) variables were not captured as part of the testing.

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

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Data analysis included: Percent agreement; and Kappa statistic to adjust for chance agreement

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 and the noise is the total variability in measured 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 evaluated by averaging over physician specific reliabilities for all providers that meet the minimum number of quality reporting events for the measure. Each provider must have at least 10 eligible reporting events to be included in this calculation.

A reliability equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability equal to one implies that all the variability is attributable to real differences in physician performance. A reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. 1

1. Adams JL, Mehrotra A, McGlynn EA, Estimating Reliability and Misclassification in Physician Profiling, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical\_reports/TR863. (Accessed on February 24, 2012.)

**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*)

Kappa statistics were calculated at the measure level for the denominator and numerator categories as a method of analyzing the reliability of measure implementation at the testing site. For analysis of the Kappa Statistics, the AMA-PCPI used Landis, J.R.; & Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics 33 (1): 159–174.

0–.20 = slight agreement

.21–.40 = fair agreement

.41–.60 = moderate agreement

.61–.80 = substantial agreement

.81–1 = almost perfect agreement

N, % agreement, kappa statistic ,(95% confidence interval)

Denominator Reliability: 120, 85.0%, 0.31 (0.10 – 0.52)

Of the 120 observations that were initially selected, 97 observations met the criteria for inclusion in the numerator analysis.

N, % agreement, kappa statistic ,(95% confidence interval)

Numerator Reliability: 97, 91%, 0.82 (0.70 – 0.93)

The reliability above the minimum level of quality reporting events was 0.99.

**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 kappa statistic value of 0.31 demonstrates fair agreement. This is due to the high observed agreement rate and the concentration of observations in the YES, YES cell (81% of all observations (97/120)). This is an example of the limitation of the Kappa statistic. While agreement can be high, if one classification category dominates, kappa can be significantly reduced. (Warrens MJ, A Formal Proof of a Paradox Associated with Cohen’s Kappa. Journal of Classification. 27:322-332, 2010; Feinstein AR, Cicchetti DV. High Agreement but Low Kappa: I. The Problems of Two Paradoxes. Journal of Clinical Epidemiology. 43:543–549, 1990)

This measure has very high reliability when evaluated above the minimum level of quality reporting 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** **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 Validity:

All PCPI performance measures are assessed for content validity by a panel of expert work group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert work group and the measures adjusted as needed. Other external review groups (eg, focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel (workgroup membership described above) was asked to 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.

Scale 1-5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5=Strongly Agree

Preventive Care and Screening: Screening for High Blood Pressure and Follow-up Documented (PQRS #317) as well as the Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan measure (PQRS #134) were chosen as suitable candidates for correlation analysis due to the similarities in patient population and domain. We hypothesize that there exists a positive association between patients aged 18 years and older who were screened for unhealthy alcohol use using a systemic screening method and who received brief counseling if identified as an unhealthy alcohol user and those who were screened for high blood pressure and a recommended follow-up plan is documented based on the current blood pressure reading as indicated. Additionally we hypothesize that there exists a positive association between patients aged 18 years and older who were screened for unhealthy alcohol use using a systemic screening method and who received brief counseling if identified as an unhealthy alcohol user and those who were screened for clinical depression using an age appropriate standardized tool and, if positive, a follow-up plan is documented on the date of the positive screen.

Providers included in the analysis met the minimum number of quality reporting events (10) and were cleaned in the same process as the PQRS dataset.

Datasets were reviewed to identify shared providers based on NPI and TIN identifiers. Correlation analysis was then performed to evaluate the association between performance scores of these shared providers.

We use the following guidance to describe correlation1:

|  |  |
| --- | --- |
| Correlation | Interpretation |
| > 0.40 | Strong |
| 0.20 - 0.40 | Moderate |
| < 0.20 | Weak |

1. Shortell T. An Introduction to Data Analysis & Presentation. Sociology 712. http://www.shortell.org/book/chap18.html. Accessed July 13, 2018.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling was positively correlated with the Preventive Care and Screening: Screening for High Blood Pressure and Follow-up Documented measure (PQRS #317) as well as the Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan measure (PQRS #134):

**PQRS #317**

Coefficient of correlation = 0.29

P-value < 0.00001

**PQRS #134**

Coefficient of correlation = 0.61

P-value < 0.00001

**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?*)

The results of the expert panel rating of the validity statement were as follows: N = 19; Mean rating = 4.32 and 84.2% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 2

3 - 1 (Neither Agree nor Disagree)

4 - 3

5 - 13 (Strongly Agree)

Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling has a moderate positive correlation and a strong positive correlation with other evidence-based process of care measures focused on preventive care services. The correlations demonstrate the criterion validity of the measure.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#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*)

Exceptions included a medical reason. Exceptions were analyzed for frequency and variability across providers.

Exceptions include:

* Documentation of medical reason(s) for not screening for unhealthy alcohol use (e.g., limited life expectancy, other medical reasons)

Exceptions were analyzed for frequency across providers.

**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*)  
Although specifications allowed for documented medical exceptions for this measure, there were no documented exceptions in this project. All sampled patients were able to be assessed.

Amongst the 8,458 physicians with the minimum (10) number of quality reporting events, there were a total of 9,785 exceptions reported. 5% of physicians reported an exception and the average number of exceptions reported by those physicians is 22.6. The proportion of exceptions to patients overall is 0.006.

**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*)

Exceptions are necessary to account for those situations when there is documentation of a medical reason for not screening for unhealthy alcohol use. Exceptions are discretionary and the methodology used for measure exception categories are not uniformly relevant across all measures; for this measure, there is a clear rationale to permit an exception for several reasons. Rather than specifying an exhaustive list of explicit reasons for exception for this measure, the measure developer relies on clinicians to link the exception with a medical reason for the decision to not screen for unhealthy alcohol use.

Some have indicated concerns with exception reporting including the potential for physicians to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by physicians, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008).

Although this methodology does not require the external reporting of more detailed exception data, the measure developer recommends that physicians document the specific reasons for exception in patients’ medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each physician’s exceptions data to identify practice patterns and opportunities for quality improvement.

Without exceptions, the performance rate would not accurately reflect the true performance of that physician. This would result in an increase in performance failures and false negatives. The additional value of increased data collection of capturing an exception greatly outweighs the reporting burden.

**References:**

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. New Engl J Med. 2008; 359: 274-84.

Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. Ann Intern Med. 2011;154:227-234.

Although the rates of exception reporting were low, exceptions are necessary to account for those situations when there is documentation of a medical reason for not screening for unhealthy alcohol use. Without exceptions, the performance rate would not accurately reflect the true performance of that physician. This would result in an increase in performance failures and false negatives.

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

**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*** [***2b4***](#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.**

Not applicable

**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**.

Not applicable

**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?

Not applicable

**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?**

Not applicable

**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.**

Not applicable

**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*)

Not applicable

*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*)**:**

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

**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*)

Not applicable

**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*)

Not applicable

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

**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)*

Data analysis performed on the measure included:

Average measure performance rate overall and by site, performance rate range by site and overall standard deviation for the measure.

Measures of central tendency, variability, and dispersion were calculated.

**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*)

Average Measure performance rate without exceptions: N= 97 Mean = 55.6% Standard Deviation= 0.4993

The performance rate by site is as follows, where n is the number of performance events by site:

1 1.000 n=40

2 1.000 n=12

3 0.044 n=45

The performance rate range is 0.956.

Based on the sample of 8,458 included physicians, the mean performance rate is 0.67 the median performance rate is 0.81 and the mode is 1.0. The standard deviation is 0.35. The range of the performance rate is 1.00, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is 0.61 (0.97–0.36).

**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?*)

Although this study captured performance on 97 events, the data

were not captured at the physician level, restricting reporting of variation in performance to the organization level only. Additionally, we are unable to present a meaningful calculation of variation in performance across organizations due to the small sample size of sites (n=3) in this study.

The range of performance from 0.00 to 1.00 suggests there’s clinically meaningful variation across physicians’ performance.

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

**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*)

This test was not performed for this measure.

**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*)

This test was not performed for this measure.

**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*)

This test was not performed for this measure.

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

**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*)

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

**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*)

This test was not performed for this measure. There was no missing data.

**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*)

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.