**National Quality Forum—Composite Measure Testing (subcriteria 2a2, 2b2-2b7, 2d)**

**Measure Number** (*if previously endorsed*)**: 2597**

**Composite Measure Title**: Substance Use Screening and Intervention Composite

**Date of Submission**: 7/7/2014

**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 specificaitions** (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 (*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). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#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)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:**  **2d1.** 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  **2d2**.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.  (*if not conducted or results not adequate, justification must be submitted and accepted)*  **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:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

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

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

Data 1

Site 1

The data are for the time period of March 2013 - October 2013.

Site 2

The data are for the time period of October 2013-December 2013.

Data 2

Site 2

The data are for the time period of March 2014 – May 2014.

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Data 1

Site 1

The data sample came from a network of community health centers across the United States serving more than 250,000 patients annually. Members in the network largely consisting of safety net organizations serving primarily low income and uninsured patients.

Site 2

The data sample came from a rural, single physician practice in the western region of the United States serving 350-500 patients a month.

Data 2

Site 2

The data sample came from a rural, single physician practice in the western region of the United States serving 350-500 patients a month.

**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*)

Data 1

Site 1

* The sample consisted of approximately 152 charts per site for a total of 32,882 eligible patients
* 2 trained abstractor reviewed the 152 patient charts
* Data collected from patients sampled in 2013
* Data abstraction performed in 2013

Site 2

* The sample consisted of approximately 110 charts per site for a total of 266 eligible patients
* 2 trained abstractors reviewed the 110 patient charts
* Data collected from patients sampled in 2013
* Data abstraction performed in 2013

Data 2

Site 2

* The sample consisted of approximately 72 charts per site for a total of 177 eligible patients
* 1 trained abstractor reviewed the 72 patient charts
* Data collected from patients sampled in 2014
* Data abstraction performed in 2014

**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**.

Not Applicable

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

**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.*

☐ **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. 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*)

* See 2b2.2 for Validity Against the Gold Standard Results

**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*)

* See 2b2.3 for Validity Against the Gold Standard Results

**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?*)

* See 2b2.4 for Validity Against the Gold Standard Results

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

**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 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)*

**Data 1 and Data 2**

**Validity Against the Gold Standard**

Data abstracted from randomly sampled patient records were used to calculate parallel forms reliability for the measure. Patient charts for abstraction were selected from visits for all patients age 18 years and older who were seen twice for any visits or who had at least one preventive care visit during the measurement period. Data analysis included:

* Percent agreement
* Kappa statistic to adjust for chance agreement

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

Data 1

Site 1 and 2

**Validity Against the Gold Standard**

This measure demonstrates almost perfect agreement in the EHR automated report in comparison to the gold standard.

*Reliability: N, % Agreement, Kappa (95% CI)*

Denominator: 801, 91%, 0.00\*\* (0.00\*\*, 0.22)

Numerator: 726, 94%, 0.89, (0.85, 0.92)

Exceptions: 726, 100%, 0.00\*\*, (0.00\*\*, 1.00)

Overall: 726, 94%, 0.89, (0.85, 0.92)

**Data 2**

Site 2

**Validity Against the Gold Standard**

This measure demonstrates substantial agreement in the EHR automated report in comparison to the gold standard.

*Reliability: N, % Agreement, Kappa (95% CI)*

Denominator: 216, 100%, Kappa is non-calculable\* (Confidence interval cannot be calculated\*, Confidence interval cannot be calculated\*)

Numerator: 216, 94%, 0.77, (0.64, 0.89)

Exceptions: 216, 99%, 0.80, (0.51, 1.00)

Overall: 216, 93.98%, 0.77, (0.64, 0.89)

\* Kappa statistics cannot be calculated because of complete agreement. Confidence intervals cannot be calculated because to do so would involve dividing by zero which cannot be done.

\*\*This is an example of the limitation of the Kappa statistic. While the agreement can be 90% or greater, if one classification category dominates, Kappa can be significantly reduced. (<http://www.ajronline.org/cgi/content/full/184/5/1391>)

**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?*)  
**Validity Against the Gold Standard**

**Data 1**

This measure demonstrates almost perfect agreement in the EHR automated report in comparison to the gold standard. 94.35% agreement was found between the abstractors and the electronic measure implemented in the EHR.

**Data 2**

This measure demonstrates substantial agreement in the EHR automated report in comparison to the gold standard. 93.98% was found between the abstractors and the electronic measure implemented in the EHR.

**Scale for interpreting kappa:**

Kappa Strength of Agreement

0.00 Poor

0.01 – 0.20 Slight

0.21 – 0.40 Fair

0.41 – 0.60 Moderate

0.61 – 0.80 Substantial

0.81 – 0.99 Almost perfect

Landis, J.R. and Koch, G.G. (1977) “The measurement of observer agreement for categorical data” in Biometrics. Vol. 33, pp. 159-174

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

**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***](#section2b4)

**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*)  
**Data 1 and Data 2**

**Validity Against the Gold Standard**

Exceptions included documentation of medical reasons for not screening for unhealthy alcohol use, tobacco use, or drug use or limited life expectancy. Exceptions were analyzed for frequency and variability across providers.

**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*)

**Data 1**

**Validity Against the Gold Standard**

* Exception Rate: 0.00%
* Comparison of the automated report to the gold standard demonstrated 100% agreement. Kappa calculations for the exceptions were 0.00\*\* Confidence interval (0.00\*\*, 1.00).

\*\*This is an example of the limitation of the Kappa statistic. While the agreement can be 90% or greater, if one classification category dominates, Kappa can be significantly reduced. (<http://www.ajronline.org/cgi/content/full/184/5/1391>)

**Data 2**

**Validity Against the Gold Standard**

* Exception Rate: 1.85%
* Comparison of the automated report to the gold standard demonstrated 99% agreement. Kappa calculations for the exceptions were 0.80 (CI 0.51, 1.00).

**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*)

Exceptions are necessary to account for those situations when it is not medically appropriate to screen for unhealthy alcohol, tobacco, and drug use. The methodology used for measure exception categories are not uniformly relevant across all measures; for each measure, there is a clear rationale to permit an exception for a medical reason or limited life expectancy. Examples have been provided in the measure exception language of instances that would constitute an exception and are intended to serve as a guide to clinicians. Rather than specifying an exhaustive list of explicit medical and limited life expectancy reasons for exception for each measure, the measure developer relies on clinicians to link the exception with a specific reason for the decision to not perform the screening required by the measure. Where examples of exceptions are included in the measure language, the measure developer has specified these reasons within the measure specifications. However this list is not intended to be an exhaustive list of reasons. Some have indicated concerns with exception reporting- 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.

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.

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

**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***](#section2b5)***.***

**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**.   
Not applicable, not an outcome or resource use measure

**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*)  
Not Applicable

**2b4.4. What were the statistical results of the analyses used to select risk factors?**Not Applicable

**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*)  
Not Applicable

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***if stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**Not Applicable

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Not Applicable

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Not Applicable

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

Not Applicable

**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?*)  
Not Applicable

\***2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not Applicable

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

**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)*   
Tobacco Use Component: We calculated interquartile range and percentiles to demonstrate statistically significant differences in performance.

2010 CMS Physician Quality Reporting Initiative:

Clinical Condition and Measure: #115 Preventive Care and Screening: Advising Smokers and Tobacco Users to Quit

# Eligible Professionals: 659,920 eligible professionals

# Professionals Reporting: >= 1 valid QDC: 12,072 professionals

% of Eligible Professionals who Reported >= 1 valid QDC: 1.8%

# of Eligible Professionals who Satisfactorily Reported: 3,388

% of Eligible Professionals who Satisfactorily Reported: 28.1%

Average Reporting Rate per Eligible Professional: 43.8%

Alcohol Use Component: We calculated interquartile range and percentiles to demonstrate statistically significant differences in performance.

2010 CMS Physician Quality Reporting Initiative:

Clinical Condition and Measure: #173 Preventive care and screening: Unhealthy Alcohol Use Screening

# Eligible Professionals: 662,161 eligible professionals

# of Professionals Reporting: >=1 valid QDC: 3,261 professionals

# Eligible Professionals who Satisfactorily Reporting: 483

% of Eligible Professionals who Satisfactorily Reporting: 14.8%

Average Reporting per Eligible Professional: 24.2%

Drug Use Component: Data analysis performed on the measure to demonstrate meaningful difference in performance score included calculating the average measure performance rate across the sites and by individual site.

**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*)

**Tobacco Use Component:**

**Physician Quality Reporting Program (PQRS) 2010**

Scores on this measure: N = 659,920 Eligible Professionals; Mean = 68.56% is the mean performance rate of NPI’s/Tax Identification Number

10th percentile: 0%

25th percentile: 32.53%

50th percentile: 98.31%

75th percentile: 100%

90th percentile: 100%

The inter-quartile range (IQR) provides a measure of the dispersion of performance and the IQR for this measure is 67.47%. A quarter of physicians had a performance score of 32.53% or less.

Confidential CMS PQRI 2010 Performance Information by Measure. Jan-Sept TAP file.

**Alcohol Use Component:**

**Physician Quality Reporting Program (PQRS) 2010**

Scores on this measure: N = 662,161 Eligible Professionals; Mean = 82.34% is the mean performance rate of NPI’s/Tax Identification Number

10th percentile: 12.50%

25th percentile: 83.64%

50th percentile: 100%

75th percentile: 100%

90th percentile: 100%

The inter-quartile range (IQR) provides a measure of the dispersion of performance and the IQR for this measure is 16.36%. A quarter of physicians had a performance score of 83.64% or less.

Confidential CMS PQRI 2010 Performance Information by Measure. Jan-Sept TAP file.

**Drug Use Component:**

**Data 1**

Site 1

N = 152

# of Eligible Patients during the Measurement Period = 32,882

Performance on the drug component measure = 13.95%

Site 2

N = 110

# of Eligible Patients during the Measurement Period = 266

Performance on the drug component measure = 74%

**Data 2**

Site 2

N = 72

# of Eligible Patients during the Measurement Period = 177

Performance on the drug component measure = 95%

**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?*)

Tobacco Use Component: PQRS data from 2010 indicates 67.47% interquartile range.

Alcohol Use Component: PQRS data from 2010 indicates 16.36% interquartile range. Despite this small range, it is important to keep in mind that PQRS data only contains data for the Medicare population, though this measure captures patients as young as 18 years old. Further, PQRS data is based on voluntary reporting with about 29% of eligible professionals participating using any reporting option in 2011.

Drug Use Component: The performance rate range is 13.95% to 95%. This demonstrates a wide range in performance. Although this study captured performance on total events, which was 33,325 for data sets 1 and 2, the data were not captured at the physician level for site 1, restricting reporting of variation in performance to the organization level only. Additionally, we are unable to present a calculation of variation in performance across organizations due to the small sample size in this study.

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

**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*)  
 Not Applicable

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
Not Applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted?*)  
Not Applicable

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

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

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

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)  
This measure was found to be reliable and feasible for implementation. All key data elements for the measure were identified during the feasibility assessment.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Not applicable

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Not applicable

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

**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.**

**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)*   
 An empirical analysis was not performed on the overall composite; rather, empirical analyses were performed on the individual composite component measures. We believe the content validity of this measure has been achieved by virtue of the noted expertise of those individuals who developed this measure. Further, we feel that this measure merits construction as a composite, as opposed to as three separate measures, due to the demonstrated topic importance of each component, documented in the individual component evidence forms. The composite construction will drive use of all three screening and intervention protocol, as a neglect of one or more protocols will result in a lower quality score.

**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*)  
Not applicable

**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)*  
Not applicable

**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**

**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)*

The approach to the composite measure algorithm is to employ a scoring methodology which identifies the number of eligible patients who received recommended care for each component measure divided by the number of eligible patients (or “opportunities”). This scoring method, known as opportunity- based scoring, is identical to that used by the Centers for Medicare and Medicaid Services (CMS) in its pay-for-performance programs.

**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*)

Not applicable

**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*)

Not applicable