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

**Measure Number** (*if previously endorsed*)**:** 0418/3148

**Measure Title**: Preventive Care and Screening: Screening for Depression and Follow-Up Plan

**Date of Submission**: 12/9/2016

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process | 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, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*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 sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment. |

|  |
| --- |
| **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 (including clinical and sociodemographic 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.   **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.  **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.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| 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*).

Part B Medicare claims data

Physician Quality Reporting System (PQRS) administrative data from registries

**Previous Submission:**

*Claim Type: Claim Carrier (B)*

*Criteria: Any HCPCS Line code in the following string: G8431, G8510, G8433, G8432, G8511*

*Additional fields requested to the standard layout: LINE\_PRCSG\_IND (included in the detail file), beneficiary name, beneficiary DOB, beneficiary DOD, beneficiary gender, beneficiary HIC, and beneficiary race*

*NPIs who had fewer than ten claims were removed from the dataset. A simple random sample of records for approximately 150 NPIs was drawn. From those 150 NPIs, a random sample of approximately 600 claims was identified. The records were then stratified by the business location address listed in the NPI registry so that the maximum number of records from each business location was limited to 10 records. This limitation was set so that the providers would not see this task as too burdensome and would be more likely to send in their records.*

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

We tested the measure using Part B Medicare claims data for encounters from 1/1/2015 to 12/31/2015.

We also tested the measure using the PQRS administrative registry data aggregated at the provider level for encounters from 1/1/2015 to 12/31/2015.

**Previous Submission:**

*Time period: 1/1/2012 – 3/31/2012*

**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.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 |

**Previous Submission:**

*N/A*

**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*)

We used Part B Medicare claims data reported by 26,169 providers, with an average of 115 patients in the denominator per provider. We used all claims with the following quality data codes (QDCs): G8431, G8432, G8433, G8510, G8511, G8940.

We used PQRS administrative registry data reported by 7,027 providers at 1,727 practices, with an average of 141 patients in the denominator per provider and an average of 550 patients in the denominator per practice.

**Previous Submission:**

*Data Sample Response Rates:*

*Number of provider requested / returned / reviewed: 155/79/77*

*Provider response rate: 51.0%*

**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*)

Total number of patients with valid denominator criteria:

Claims data: 3,002,169

Registry data: 989,092

**Descriptive characteristics of patients – claims data:**

Sex

Female: 1,759,168 (58.6%)

Male: 1,243,001 (41.4%)

Race

Asian: 30,234 (1.0%)

Black: 363,378 (12.1%)

Hispanic: 57,189 (1.9%)

Native American: 10,992 (0.4%)

White: 2,472,318 (82.4%)

Other: 36,502 (1.2%)

Unknown: 31,556 (1.1%)

Age

12-17: 62 (0%)

18-64: 589,536 (19.6%)

65+: 2,412,571 (80.4%)

PQRS administrative data from registries did not include descriptive patient data.

**Previous Submission:**

*Data Sample Response Rates:*

*Number of records requested / returned / reviewed: 641/294/275*

*Provider response rate 45.9%*

**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**.

We used both sources (claims and registry) to assess providers’ measure performance and provider-level reliability. We also used registry data to assess practice-level measure performance and reliability. We examined demographics and disparities in claims data but not registry data because registry data do not include patient characteristics such as age, sex, or race.

**Previous Submission:**

*N/A*

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We aggregated performance scores in claims data by race, sex, and age to look for disparities. Claims data do not include information about income or other sociodemographic information. Registry data are aggregated at the provider-level and do not include sociodemographic variables.

**Previous Submission:**

*N/A*

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

**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*)

**Previous Submission:**

*[Critical data elements]*

**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*)

We calculated reliability using a widely accepted method that is outlined in J.L. Adams’ (2009) technical report titled “The Reliability of Provider Profiling: A Tutorial.” In this context, reliability represents a measure’s ability to confidently distinguish the performance of one physician from another. As discussed in the report, “Conceptually, [this method assesses] the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error.” In this method, reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within providers), whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities. Although, there is not a clear cut-off for minimum reliability level, values above 0.7 are considered sufficient to see differences between physicians (or practices) and the mean, and values above 0.9 are considered sufficient to see differences between individual physicians or practices.

Adams, J.L. (2009). *The reliability of provider profiling: A tutorial* (TR-653-NCQA). Santa Monica, CA: RAND Corporation. Retrieved November 14, 2016, from <http://www.rand.org/pubs/technical_reports/TR653.html>

**Previous Submission:**

*Crude agreement rates were calculated along with prevalence adjusted kappa (PAK), Cohen´s kappa values and corresponding confidence intervals. Cohen´s kappa represents chance-corrected proportional agreement. High prevalence of responses in a small number of cells is known to produce unexpected results known as the "kappa paradox" When the prevalence of a rating in the population is very high or low, which was noted in the testing of this measure, the value of kappa may indicate poor reliability even with a high observed proportion of agreement. In such cases, as with this measure, PAK is shown to provide an additional interpretation of agreement when the prevalence of responses is concentrated in a small number of cells.*

**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*)

**Performance measure score reliability:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data source | Number of providers/ practices | Between-provider variance | Reliability mean | Reliability median | Reliability standard deviation | Reliability min/max |
| Claims | 26,169 | .21 | .99 | 1.0 | .03 | .62 - 1.0 |
| Registry – provider level | 7,027 | .19 | .99 | 1.0 | .04 | .60 - 1.0 |
| Registry – practice level | 1,797 | .18 | .99 | 1.0 | .05 | .59 - 1.0 |

**Previous Submission:**

*Overall Reliability: Claims vs. Independent Review:*

*Numerator: 79.2% agreement, PAK=.60 (CI .51 -.69), Kappa=.38(CI .25 - .50)*

*Denominator Exclusions: 93.0% agreement, PAK=.86 (CI .80 - .92), Kappa .64 (CI .49 - .79)*

*Valid Denominator Criteria: 100.0%*

*Of the 275 total cases reviewed, 240 cases were not reported as exclusions and were reviewed for numerator agreement as compared with the code reported on the claim.*

*In 191 of 240 (79.0%) cases, reviewers agreed with whether or not the case met the numerator criteria based on the code submitted with the claim. A large proportion of those cases fell into the ”met numerator criteria” category. Of the 191 cases where agreement was present, 165 cases(86.4%) agreed the case met the numerator criteria while 26 cases(13.6%) agreed the case did not meet the numerator criteria. This prevalence should be considered in the interpretation of kappa scores.*

*Agreement with claims on denominator exclusions was high, with 93% agreement and prevalence adjusted kappa of .86 (95% CI .80 – .92).*

*Inter-Rater Reliability: Quality Insights of Pennsylvania Internal RN Reviewer (QIP) vs. Independent External RN Reviewer (ALPS):*

*Numerator: 89.7% agreement, PAK=.80 (CI .70 -.89), Kappa=.75 (CI .64 - .86)*

*Denominator Exclusions: 66.5% agreement, PAK=.39 (CI .30 -.48), Kappa .18 (CI .09 -.27)*

*Valid Denominator Criteria: 100%*

*All records without valid denominator criteria were removed prior to reliability assessment. Denominator agreement was 100%.*

*Reporting of this measure demonstrates high numerator reliability. Disagreements between ALPS and claims were identified when the ALPS abstractors was unable to find the documentation of the Age Appropriate Standardized Screening Tool used to perform the depression screening. The 2012 PQRS specification does not directly instruct the provider to document the tool used. Updating the measure specifications to add specific guidance to the provider to include documentation of the tool used in the patient’s medical record could help to improve the reliability of this measure. Changes are in the process of being modified for the measure specification which will be finalized for the 2014 reporting year.*

**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?*)

With average reliability scores of 0.99 for both claims and registry reported data, this measure demonstrates a high level of reliability to detect real difference in performance scores.

**Previous Submission:**

*Results of inter-rater reliability for two independent reviewers reflect differences in interpreting the exclusion criteria. Follow-up debriefings determined the main reason for mismatch was the use of different data sources to determine eligibility for exclusion. Upon review of the 2012 PQRS specification, the exclusion criterion was found to lack clarity which would have minimized the variability of interpretation when identifying exclusions. Additionally, while re-tooling this measure for an EHR specification in early 2012, this lack of specificity of the exclusion criterion was identified by Quality Insights of Pennsylvania (QIP) e-specification team. As a result, the 2013 PQRS Claims and Registry specification was updated and the exclusion criteria were changed to add more clarity. Additionally, the 2014 EHR specifications were re-tooled with updated exclusion language replacing the previous exclusions “Patient was referred with a diagnosis of depression” and “Patient has been participating in on-going treatment with screening of clinical depression in a preceding reporting period” with “Patient has an active diagnosis of depression or bipolar disease.” Guidance will be added to the all specifications to define “active diagnosis” to reduce variability in this definition.*

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

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**Previous Submission:**

*[Face validity]*

**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)*

We surveyed 12 clinicians eligible to report this measure—none of whom advised on measure development—to rate face validity. We provided measure specifications and asked them to rate their agreement with the following statement: “The performance 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 rating scale offered five options: 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; 5 = strongly agree.

**Previous Submission:**

*Quality Insights conducts an annual environmental scan to evaluate the most current research and evidence-based guidelines. A Technical Expert Panel (TEP), composed of subject matter specialists and experts with technical measure expertise, evaluates the results of the scan and provides recommendations based on the scientific merits of the evidence using the Strength of Recommendation Taxonomy (SORT). The TEP also reviews and establishes the measure’s capability to capture what it is designed to capture using a consensus process.*

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

1 – Strongly disagree – 0 votes

2 – Disagree – 3 votes

3 – Neither agree nor disagree – 0 votes

4 – Agree – 6 votes

5 – Strongly agree – 3 votes

**Previous Submission:**

*N/A*

**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?*)

Nine of 12 experts (75 percent) agreed or strongly agreed that the measure accurately reflects quality. Experts who disagreed raised concerns related to patient compliance, documentation burden, and a preference that the measure specify one screening tool for adolescents, rather than several tools from which providers can choose.

**Previous Submission:**

*Face validity is established by subject matter specialists and experts who determine that the measure represents the process of interest.*

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

**2b3. EXCLUSIONS ANALYSIS**

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

**Previous Submission:**

*N/A*

**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*)

We assessed the frequency of exclusions using claims and registry data from 2015.

**Previous Submission:***N/A*

**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*)

In 2015 Medicare claims, 3.6 percent of eligible encounters qualified as exclusions. In 2015 registry data, 4.9 percent of eligible encounters were reported as exclusions.

**Previous Submission:**

*N/A*

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The low rates suggest that the exclusions will not unduly distort measure performance. Although they may not dramatically change measure performance, exclusions allow for provider discretion in determining whether to screen patients for depression and provide follow-up interventions, improving the measure’s face validity.

**Previous Submission:**

*N/A*

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

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b4.1.1 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.**

N/A

**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**.

N/A

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

N/A

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

N/A

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)**

N/A

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)  
*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

N/A

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

N/A

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

N/A

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

N/A

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

N/A

**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*)

N/A

**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*)

N/A

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

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

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We used claims and registry data from calendar year 2015 to calculate measure performance scores and assess the distribution of performance using statistical measures of central tendency (mean and median), variation (standard deviation), and spread (interquartile range and rates by percentile).

Using claims data, we calculated chi-square statistics to test for significant differences between expected and observed aggregate performance scores based on patients’ race, sex, and age. Registry data do not include descriptive patient data.

**Previous Submission:**

*Dates of service from 1/1/2012 to 3/31/2012*

*Total Claims Submitted with any G code (G8431, G8510, G8433, G8432, G8511): 10,004*

*Valid Denominator Criteria: 7709 (77.1% of total)*

*Performance Exclusion: 1126 (14.6% of valid submissions)*

*Total tested claims sampled and reviewed: 275 records from 77 providers*

*Valid denominator criteria: 275/275 (100.0% of total)*

*Sample Performance Exclusion (claims based): 35 (12.7% of valid)*

*Aggregate measure performance rate for sample (claims based): 216/275 (78.5%)*

*Aggregate and provider (NPI) performance rates were calculated from Part B claims with dates of service from 1/1/2012 through 3/3/2012. Data from the testing sample were analyzed at the provider level. Performance rates are derived from G codes submitted for the Physician Quality Reporting System (formerly PQRI). Code submissions are voluntary and providers who report may not be representative of all eligible professionals. Performance rates cannot be generalized to the population.*

**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*)

**Distribution of performance scores**

Registry data (Provider-level)

* N: 7,027
* Mean: 50.7%
* STD: 44.3%
* Interquartile range: 99.7%
* 10th percentile: 0.0%
* 20th percentile: 0.0%
* 30th percentile:2.2%
* 40th percentile: 17.8%
* Median: 50.8%
* 60th percentile: 85.7%
* 70th percentile: 100.0%
* 80th percentile: 100.0%
* 90th percentile: 100.0%

Registry data (Practice-level)

* N: 1,797
* Mean: 55.2%
* STD: 41.8%
* Interquartile range: 95.0%
* 10th percentile: 0.0%
* 20th percentile: 1.2%
* 30th percentile: 12.1%
* 40th percentile: 39.9%
* Median: 63.4%
* 60th percentile: 85.7%
* 70th percentile: 99.2%
* 80th percentile: 100.0%
* 90th percentile: 100.0%

Claims data (Provider-level)

* N: 26,169
* Mean: 63.8%
* STD: 45.9%
* Interquartile range: 100.0%
* 10th percentile: 0.0%
* 20th percentile: 0.0%
* 30th percentile: 5.9%
* 40th percentile: 85.7%
* Median: 100.0%
* 60th percentile: 100.0%
* 70th percentile: 100.0%
* 80th percentile: 100.0%
* 90th percentile: 100.0%

**Performance results by population groups, claims data:**

Age groups  
18–64: 35.7%

65+: 36.7%

(*X2* = 207.5; *df*: 1; *N*: 3,002,107; *p* < 0.0001)

(Age category 12–17 was excluded due to small sample size)

Race

Asian: 58.4%

Black: 26.8%

Hispanic: 43.2%

Native American: 73.9%

White: 37.1%

Other: 50.0%

Unknown: 38.9%

(*X2* = 31,993; *df*: 6; *N*: 3,002,169; *p* < 0.0001)

Sex  
Female: 37.6%

Male: 34.8%

(*X2*: 2,575.2; *df*:1; *N*: 3,002,169; *p* < 0.0001)

**Previous Submission:**

*Performance rates were calculated at the provider level for claims from the time period 1/1/2012 to 3/31/2012.*

*Aggregate measure performance rate: 5463/6583 (83.0%)*

*Distribution of provider scores (by NPI): N=459, Mean = 84.3%, Median=100.0%, SD=.339 Range=100*

*10th percentile: 0%, 25th percentile: 100.0%; 75th percentile 100.0%*

*Testing sample*

*Measure performance rate (claims based): 216/275 (78.5%)*

**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?*)

Reported performance rates indicate a wide degree of variation in both claims and registry data, and many clinicians have the potential to improve the rates of depression screening and follow-up. The differences in performance rates by age and sex were statistically significant but small in magnitude and therefore of limited clinical significance. Among racial groups, clinically significant differences are apparent, and quality improvement efforts should attempt to address these disparities. However, we did not stratify the measure based on race because: (1) many other process measures show similar racial disparities and (2) stratifying the measure would significantly complicate implementation, reporting, and interpretation. Providers report this measure voluntarily, and reported performance rates may not represent the total eligible population.

**Previous Submission:**

*N/A*

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

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

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

**Note***: This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

Claims/registry and electronic clinical quality measure (eCQM) specifications are aligned across reporting methods. As directed by NQF, eCQM testing data are submitted separately as NQF 3132.

**Previous Submission:**

*N/A*

**2b6.1. 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*)

Entities report the measure using a single data source. We did not compare performance rates between the claims/registry measure and the eCQM because the eCQM is submitted separately. However, we designed the specifications for all the data sources to maximize alignment and consistency.

**Previous Submission:**

*N/A*

**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*)

N/A

**Previous Submission:**

*N/A*

**2b6.3. 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*)

N/A

**Previous Submission:**

*N/A*

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

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

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)**

There were no missing data in our analysis because our testing relied on all of the cases reported for this measure via claims and registries for the PQRS program from 1/1/2015 to 12/31/2015. In claims data, if an encounter record included a relevant CPT encounter code and QDC, we assumed it was eligible for the measure, and we used QDCs to group encounters into performance categories (met performance, failed performance, excluded). If these data were not available, we did not include the encounter in our analysis. Our registry data source included provider-level results for all providers who participated in registry reporting to the PQRS program. These data showed the number of patients per provider who met performance, failed performance, and were excluded.

**Previous Submission:**

*N/A*

**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*)

There were no missing data in our claims or registry analyses. As noted above, claims lacking QDCs were assumed to be not relevant and excluded from our analysis.

**Previous Submission:**

*N/A*

**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*)

We used data from all eligible professionals who reported the measure using claims or registry data, which meant there were no systematic missing data. Because reporting is voluntary, the reporting population is not necessarily representative of the total eligible population and results are not generalizable to the overall eligible population. According to the appendix tables in the 2014 PQRS Reporting Experience, 7.5 percent of eligible professionals reported the measure in 2014 (Centers for Medicare & Medicaid Services, 2016).

Centers for Medicare & Medicaid Services. (2016, April). *2014 PQRS reporting experience, including trends (2007-2015).* Retrieved November 14, 2016, from https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

**Previous Submission:**

*N/A*