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

**Measure Number** (*if previously endorsed*)**:** 1932

**Measure Title**: Diabetes screening for people with schizophrenia or bipolar disorder who are prescribed antipsychotic medications (SSD)

**Date of Submission**: 4/2/2018

**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*).

**2018 Submission**

N/A

**2012 Submission**

N/A

**1.3. What are the dates of the data used in testing**? 2018 submission: 2016 data; 2012 submission: 2007 data

**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*)

**2018 Submission**

Population for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 202 Medicaid plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Population for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 145 Medicaid health plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

**2012 Submission**

Using Medicaid Analytic Extract (MAX) claims data from 2007 we included Medicaid recipients from 22 states who met the following criteria (1) enrolled in fee-for-service plans\* (2) disability as the basis of eligibility; and (3) continuously enrolled in Medicaid for 10 months.

The data came from the following states: Alabama, Alaska, California, Connecticut, DC, Georgia, Idaho, Illinois, Indiana, Iowa, Louisiana, Maryland, Missouri, Mississippi, Nevada, New Hampshire, North Carolina, North Dakota, Oklahoma, South Dakota, West Virginia and Wyoming.

**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*)   
**2018 Submission**

Patient sample for measure score reliability testing: In 2016, HEDIS measures covered 47 million Medicaid members. Data are summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median number of eligible patients per plan |
| Medicaid | 202 | 1,018 |

Beneficiary Sample for Construct Validity Testing: In 2016, HEDIS measures covered 47 million Medicaid beneficiaries. Data is summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of plans | Median number of eligible patients per plan |
| Medicaid | 202 | 1,018 |

**2012 Submission**

We drew two analytic samples from the beneficiaries. Beneficiaries who had a primary diagnosis of schizophrenia on either one inpatient or two outpatient claims on different days were included in our schizophrenia sample. We also tested beneficiaries who had a primary diagnosis of either schizophrenia or bipolar disorder on either one inpatient or two outpatient claims on different days. Overall, there were 98,412 beneficiaries in the schizophrenia sample and 130,529 beneficiaries in the schizophrenia or bipolar disorder sample.

Beneficiaries ranged in age from 25 – 64 years. Just under half of the schizophrenia population was female (49.2%) while nearly 55% of beneficiaries with schizophrenia or bipolar disorder were female (54.8%). About 7% of both samples were Hispanic and African-Americans comprised 34% and 39%, respectively, of the schizophrenia or bipolar disorder and schizophrenia samples.

(\*Beneficiaries enrolled in managed care plans (e.g. BHO or HMO plans) that provided usable claims records were included. About 1% of the schizophrenia sample was enrolled in a BHO (1.4%) and 11.5% were enrolled in an HMO).

**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**.

**2018 Submission**

N/A

**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.

**2018 Submission**

We did not analyze performance by social risk factors.

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

**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*)

**2018 Submission**

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´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. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is 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 performance. 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 performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

**2012 Submission**

The relevant unit of analysis for the proposed measures is aggregated state-level performance. Therefore, we conducted an analysis of test-retest reliability for state results to assess the reliability of state-level performance. To assess stability of state-level performance over time, we computed quartiles of performance based on the state distribution for each measure and assigned each state a score reflecting each state’s performance relative to other states in the distribution during the measurement year. For example, a state in the top quartile of all states in 2007 for a given measure would be assigned a performance quartile score of ‘1’ for 2007. This method was replicated for each measure. Next, we repeated this method using 2008 claims data and examined stability of performance quartile between 2007 and 2008.

We also report Pearson correlations measuring the association between 2007 and 2008 measure performance for the 16 states with data.

**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*)  
**2018 Submission**

Beta-Binomial Statistic:

|  |
| --- |
| Medicaid |
| 0.959 |

**2012 Submission**

Overall, 4 of 16 states (25%) had no change in performance quartile between 2007 and 2008. State performance for this measure correlated at r=0.33. In general, the measure showed good test-retest reliability. The result also indicated that 2007 performance on this measure accounted for 11% of the variance in 2008 scores.

**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?*)  
Interpretation of measure score reliability testing: The testing suggests the measure has strong reliability with beta binomial result of 0.959 exceeding the 0.7 threshold.

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

**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)*  
**2018 Submission**

We assessed construct and face validity for this measure.

Method of testing construct validity: We tested for construct validity by exploring whether the Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications measure is correlated with the Diabetes Monitoring for People With Diabetes and Schizophrenia measure. We hypothesized that organizations that perform well on Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications should perform well on the Diabetes Monitoring for People With Diabetes and Schizophrenia measure because the two measures both focus on patients with schizophrenia and whether they received care for diabetes.

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable

Method of Assessing Face Validity: We describe below NCQA’s process for both measure development, and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA’s Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assures that measures we use are valid.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs – whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. On average, NCQA receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQA’s Board of Directors will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported, and audited before it is used for public accountability or accreditation. This is not testing – the measure was already tested as part of its development – rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA’s Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year’s data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year’s HEDIS Volume 2.

**2012 Submission**

Validity was assessed using several complementary methods.

Face validity was assessed through a multistakeholder Technical Advisory Group responsible for overseeing measure development. Additionally, face validity was captured through a public comment period and a series of focus groups involving the Medicaid Medical Directors Learning Network, Managed Behavioral Health Care Organizations, and State Mental Health Commissioners and Medical Directors. The panelists assessed the usability and feasibility of the measures.

Concurrent validity was assessed via Medicaid resource utilization from the Medicaid claims data. We examined rates of schizophrenia-related hospital and emergency room utilization as well as total Medicaid costs comparing beneficiaries in the highest and lowest performance quartiles for each measure.

Convergent and discriminant validity were assessed using the Medicaid Analytic Extract (MAX) from Medicaid claims in using 2007 data. Pearson correlation coefficients were used to assess measure correlations. We hypothesized similar measures (e.g. screening and monitoring) would be correlated and (b) process measures would have negative correlations with measures of adverse events (e.g. mental health emergency room utilization).

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
**2018 Submission**

Statistical results of construct validity testing: The results in Table 1 indicate that there is a statistically significant (P<0.05) and positive relationship between the Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications measure and the Diabetes Monitoring for People With Diabetes and Schizophrenia measure.

**Table 1. Correlations in Medicaid Measures – 2016**

|  |  |
| --- | --- |
|  | **Pearson Correlation Coefficient** |
| Diabetes monitoring for people with diabetes and schizophrenia |
| Diabetes screening for people with schizophrenia or bipolar disorder who are using antipsychotic medications | 0.25 |

Note: p<0.05

Results of face validity assessment: Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity.

**2012 Submission:**

Face validity:

The measures were deemed important, usable, and feasible to collect by the Technical Advisory Group overseeing the measure development, as well as focus groups with the Medicaid Medical Directors Learning Network, Managed Behavioral Healthcare Organizations, and State Mental Health Commissioners and Medical Directors.

Among 22 states, the measure had a minimum value of 2.3%, mean=12.1%, 25th percentile=8.4%, median=10.3%, 75th percentile=16.7% and a maximum value of 28.2%.

Concurrent validity:

Beneficiaries in the lowest performing states for the measure had higher rates of schizophrenia related hospitalization and ED use (24.3% and 26.6%, respectively) than individuals in the highest performing states (18.1% and 24.5%, respectively).

Concurrent and discriminant validity:

Performance on the measure was significantly correlated with cardiovascular screening (r=.276, p<.001).

**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?*)  
**2018 Submission**

Interpretation of construct validity testing: The two measures had statistically significant positive correlation, which indicates the measure has good construct validity.

Interpretation of systematic assessment of face validity: NCQA’s expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Diabetes screening for people with schizophrenia or bipolar disorder who are prescribed antipsychotic medications (SSD)* is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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

**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*)  
Testing was not performed for exclusions.

**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*)  
Testing was not performed for exclusions.

**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*)  
Testing was not performed for exclusions.

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

**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.**

**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**.

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

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

**2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

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

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

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

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

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*   
**2018 Submission**

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans’ performance is significantly different from each other.

**2012 Submission**

Pearson correlations, means and percentiles are reported.

**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*)  
**2018 Submission**   
HEDIS 2017 Variation in Performance across Health Plans

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Medicaid | 1,464 | 80.7 | 5.8 | 74.0 | 77.5 | 81.0 | 84.2 | 87.4 | 6.7 | <0.001 |

EP: Eligible Population, the average denominator size across plans submitting to HEDIS

IQR: Interquartile range

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

**2012 Submission**

Among 22 states, the measure had a minimum value of 2.3%, mean=12.1%, 25th percentile=8.4%, median=10.3%, 75th percentile=16.7% and a maximum value of 28.2%.

**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?*)  
**2018 Submission**

The difference between the 25th and 75th percentile is statistically significant for the Medicaid product line. For Medicaid plans, there is a 6.7 percentage point gap between 25th and 75th percentile plans. This gap represents an average 98 more patients with schizophrenia or bipolar disorder who were dispensed an antipsychotic medication having diabetes screening test during the measurement year in high performing Medicaid plans compared to low performing plans (estimated from average health plan eligible population).

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

**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*)

**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*)

**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*)

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

**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*)  
**2018 Submission**

This measure is collected with a complete sample.

**2012 Submission**

There is no bias on this measure due to missing data.

**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*)

**2018 Submission**

This measure is collected with a complete sample.

**2012 Submission**

There is no bias on this measure due to 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*)

**2018 Submission**

This measure is collected with a complete sample.

**2012 Submission**

There is no bias on this measure due to missing data.