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

**Measure Number** (*if previously endorsed*)**: 2601 (New Measure)**

**Measure Title**: **Body Mass Index Screening and Follow Up for People with Serious Mental Illness**

**Date of Submission**: **7/25/2014**

**Type of Measure:**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | ☐ Outcome (*including PRO-PM*) |
| ☐ 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). |
|  |

|  |
| --- |
| **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.  **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 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*).

**Not applicable.**

**1.3. What are the dates of the data used in testing**? **2011-2012**

**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 |

**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*)

**RELIABILITY AND MEANINGFUL DIFFERENCES**

**Three health plans were included in testing and analysis. The plans consisted of a Medicaid plan for non-disabled adults with enrollment of approximately 130,000 patients, a Special Needs Plan for dual-eligible patients (Medicare and Medicaid) with enrollment of approximately 13,000 patients, and a Medicaid plan for disabled adults with enrollment of approximately 13,000 patients. The plans were geographically dispersed and included plans from the West, Midwest, and the East regions of the US.**

**FACE VALIDITY**

**This measure was tested for validity with an expert panel (n=16), focus group (n=29), and public comment (n=20).**

**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*)

**The sample was identified using administrative/claims data for individuals enrolled at each health plan from January 1, 2011 – December 31, 2012. The testing for this measure was conducted in conjunction with testing for measures addressing diabetes and hypertension. Therefore, we instructed plans to draw random samples of three patient groups: patients with Serious Mental Illness (SMI) only, patients with SMI and Diabetes, and patients with SMI and Hypertension. Plans used claims data to select the samples. For all sampled patients, we instructed plans to review patient medical and behavioral health records. Health plan staff abstracted the medical and behavioral health records separately.**

**The final sample included 1,014 records, of which 306 patients were SMI only, 258 were SMI and Diabetes, and 450 were SMI and Hypertension (Table 1). Due to overlap of patients with diabetes and hypertension, the testing included 884 unique patients across the three health plans.**

**Across the 3 SMI groups, 62-71% had a medical record available for review (Table 1) and 20-28% did not have an ambulatory visit and therefore did not have a record for review. The health plan was not able to obtain an ambulatory record for 7-11% of patients who had an ambulatory visit during the measurement year. If an organization cannot find the medical record, the patient remains in the measure denominator and is considered noncompliant for the numerator.**

**Fewer patients had a behavioral health record for review because either they did not have a behavioral health visit or the health plan was unable to access a behavioral health record (Table 2). Across the 3 patient groups, 23-32% had a behavioral health record available for review; 36-47% did not have a behavioral health visit and therefore did not have a record for review; and 21-33% had a behavioral health visit during the measurement period but did not have a behavioral health record available for review (Table 2).**

**Overall, across the 3 patient groups, 67-77% had either a medical or behavioral health record available for review (Table 3); 15-21% did not have either type of visit and therefore did not have a record for review, and 8-12% had at least one of either type of visit but the health plan was unable to obtain a record.**

**Table 1. Number of Patient Records for Review: Medical Records**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **SMI only** | **SMI & Diabetes** | **SMI & Hypertension** |
| **Patients Sampled** | 306 | 258 | 450 |
| **a. Patients with a medical record for review** | 216 (70.6%) | 183 (70.9%) | 277 (61.6%) |
| **b. Patients who did not have a medical visit and therefore did not have a record** | 69 (22.5%) | 52 (20.2%) | 125 (27.8%) |
| **c. Patients who had a medical visit but did not have a record available for review** | 21 (6.7%) | 23 (8.9%) | 48 (10.7%) |

**Table 2. Number of Patient Records for Review: Behavioral Health Records**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **SMI only** | **SMI & Diabetes** | **SMI & Hypertension** |
| **Patients Sampled** | 306 | 258 | 450 |
| **a. Patients with a behavioral health record for review** | 95 (31.0%) | 59 (22.9%) | 145 (32.2%) |
| **b. Patients who did not have a behavioral health visit and therefore did not have a record** | 109 (35.6%) | 116 (45.0%) | 212 (47.1%) |
| **c. Patients who had a behavioral health visit but did not have a record available for review** | 102 (33.3%) | 83 (32.2%) | 93 (20.7%) |

**Table 3. Number of Patient Records for Review: Medical or Behavioral Health Records**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **SMI only** | **SMI & Diabetes** | **SMI & Hypertension** |
| **Patients Sampled** | 306 | 258 | 450 |
| **a. Patients with either medical or behavioral health record for review** | 235 (76.8%) | 190 (73.6%) | 302 (67.1%) |
| **b. Patients who did not have any visit and therefore did not have a record** | 47 (15.4%) | 41 (15.9%) | 96 (21.3%) |
| **c. Patients who had either a medical or behavioral health visit but did not have either record available for review** | 24 (7.8%) | 27(10.5%) | 52 (11.6%) |

**Across plans, there was diversity in diagnosis, age, and gender. We instructed health plans to attempt to sample equal numbers of patients with each serious mental illness diagnosis (depression, bipolar I, and schizophrenia). However, due to variation in the prevalence of diagnosis across groups and plans, some plans were unable to sample an equal number of people with each diagnosis (see Table 4).**

**Notably, the Medicaid disabled plan had fewer individuals with bipolar I disorder. As expected, the Dual SNP had a larger share of adults over age 50. The Dual SNP plan had more females; the Medicaid Disabled plan had more males, and the Medicaid Adult plan had a more even gender mix (except for more male patients in the SMI-only group).**

**Table 4. Distribution of Diagnosis, Age, and Gender by Patient Group and Plan**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **SMI only** | | | | **SMI + Hypertension** | | | | **SMI +Diabetes** | | | |
|  | **All Plan (%)** | **Dual SNP (%)** | **Medicaid Disabled (%)** | **Medicaid Adult (%)** | **All Plan (%)** | **Dual SNP (%)** | **Medicaid Disabled (%)** | **Medicaid Adult (%)** | **All Plan (%)** | **Dual SNP (%)** | **Medicaid Disabled (%)** | **Medicaid Adult (%)** |
| **Patients Sampled (N)** | 306 | 102 | 102 | 102 | 450 | 135 | 112 | 203 | 258 | 130 | 40 | 83 |
| **Diagnosis** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Major depression** | 30 | 33 | 40 | 16 | 36 | 33 | 56 | 26 | 33 | 33 | 53 | 24 |
| **Schizophrenia** | 37 | 33 | 38 | 40 | 33 | 33 | 35 | 33 | 36 | 33 | 38 | 41 |
| **Bipolar I disorder** | 33 | 33 | 22 | 44 | 31 | 33 | 9 | 42 | 30 | 33 | 10 | 35 |
| **Age** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Age 18-50** | 69 | 54 | 71 | 82 | 54 | 21 | 43 | 81 | 42 | 27 | 35 | 69 |
| **Age >50** | 31 | 46 | 29 | 18 | 46 | 79 | 57 | 19 | 59 | 73 | 65 | 31 |
| **Gender** |  |  |  |  |  |  |  |  |  |  |  |  |
| **Male** | 53 | 39 | 56 | 63 | 47 | 38 | 59 | 46 | 45 | 39 | 50 | 52 |
| **Female** | 47 | 61 | 44 | 37 | 53 | 62 | 41 | 54 | 55 | 61 | 50 | 48 |

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

**The full sample of 884 unique patients was used to examine meaningful differences in performance. A random subsample of 223 patient medical records was double-abstracted and used for inter-rater reliability testing.**

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

**2a2. RELIABILITY TESTING**

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

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

**Reliability was tested by assessing inter-rater reliability. Inter-rater reliability assesses whether two abstractors, reviewing the same data from the same data source, agreed on whether the patient met the requirements for the numerator, denominator, or exclusions for the measure. Inter-rater reliability was calculated based on data collected by two raters on 223 patient medical records randomly selected across the 3 plans. We used the kappa statistic, a measure of agreement adjusted for chance to quantify agreement.**

**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*)

**The table below shows the number of patients evaluated for agreement, the percent agreement, and the Kappa statistics (with its 95% confidence interval).**

**Table 5. Percentage of Agreement and Kappa Statistic**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **All Plans, All SMI Patients** | | | |
|  | **Number of patients evaluated for agreement** | **% Agreement** | **Kappa** | **95% C.I.** |
| **Denominator before exclusions (Not applicable because denominator is identified using claims data)** | NA | NA | NA | NA |
| **Exclusions (Calculated among patients found by both abstractors to qualify for the denominator before exclusions)** | 223 | 100.0 | 1.00 | 1.00, 1.00 |
| **Numerator before exclusions (Calculated among patients found by both abstractors to qualify for the denominator before exclusions)** | 223 | 92.8 | 0.85 | 0.78, 0.92 |
| **Overall: Numerator after exclusions (Calculated among patients found by both abstractors to qualify for the denominator after exclusions)** | 216 | 92.6 | 0.84 | 0.77, 0.92 |

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

**The kappa results showed almost perfect inter-rater reliability for the numerator, exclusion, and the overall measure performance. The denominator was identified using claims data and did not require chart abstraction or assessment of inter-rater reliability. This indicates that the measure results can be reliably collected and reported.**

**For reference, the Kappa statistic has the following interpretation (Landis & Koch, 1977):**

**Table 6. Interpretation of Kappa**

|  |  |
| --- | --- |
| **Kappa** | **Interpretation** |
| 0.00 | Poor Agreement |
| 0.01 – 0.20 | Slight Agreement |
| 0.21 – 0.40 | Fair Agreement |
| 0.41 – 0.60 | Moderate Agreement |
| 0.61 – 0.80 | Substantial Agreement |
| 0.81 – 0.99 | Almost Perfect Agreement |

**Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-174.**

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

**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*)

**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)*

**CRITICAL DATA ELEMENT VALIDITY TESTING**

**Good inter-rater reliability for data elements also supports the data element validity of the measure. For reliability testing, two reviewers used the same measure specification and drew similar conclusions from the same “gold-standard” data source (medical record). This testing demonstrated that two independent reviewers looking at the same full medical record had high agreement on every data element and the overall performance measure score.  We believe this testing demonstrates not only reliability but also validity, that is to say the accuracy of the measure specification to identify all data elements from the medical record.  The steps for testing inter-rater reliability were described in section 2a2.**

**SYSTEMATIC ASSESSMENT OF FACE VALIDITY**

**Our field test addressed the face validity of the measure specification by several types of stakeholder input.**

**A multistakeholder technical expert panel of 16 individuals consisting of health plan representatives, consumers, state Medicaid and behavioral health representatives, and quality measurement experts was convened and provided input throughout the measure development process, including review of the field test results and recommendations for final specifications.**

**In addition, four multistakeholder focus groups that included 29 representatives from Medicaid plans, state Medicaid, mental health and substance abuse agencies, integrated care systems, consumers/advocates, and other health care organizations reviewed and commented on the draft specifications and field test results.**

**We also received feedback from a two-week public comment period hosted on NCQA’s online public comment system. The public comment notification was submitted to stakeholders representing consumers, health plans, clinicians, quality measurement and behavioral health experts.**

**ICD-10 CONVERSION**

**The goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.**

***Steps in ICD-9 to ICD-10 Conversion Process***

1. **NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.**
2. **NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.**
3. **NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.**
4. **As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.**
5. **Post ICD-10 code recommendations for public review and comment.**
6. **Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.**
7. **NCQA staff finalize ICD-10 code recommendations.**

***Tools Used to Identify/Map to ICD-10***

**All tools used for mapping/code identification from CMS ICD-10 website (**[**http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html**](http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html)**).**

**GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.**

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

**CRITICAL DATA ELEMENT VALIDITY TESTING**

**High inter-rater reliability supports the accuracy of the measure specifications.**

**SYSTEMATIC ASSESSMENT OF FACE VALIDITY**

**Participants in the technical expert panel and all four of the stakeholder focus groups supported moving forward with this measure. Out of 14 total comments that were received from public comment on this measure, 8 (57%) supported or supported the measure with modifications.** **Other commenters did not support the measure due to concerns about the burden of record review but did not express concern about validity of the measure. There were concerns that some plans may not cover certain services such as nutrition counseling unless the patient is morbidly obese or has diabetes. The TEP expressed concern over the burden of chart reviews but felt the importance, usability, and validity of the measure outweighed the concerns about burden.**

**The TEP also recommended changing the numerator requirement to include two follow up events (e.g., counseling). This change raises the intensity of service to address the high risk status of the SMI population and to take advantage of health plans’ opportunity/responsibility for follow-up care beyond the visit. This adaptation strengthened the face validity of the measure.**

**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?*)

**The stakeholder input suggests that this measure is valid for assessing effectiveness of BMI screening and follow up among people with SMI. The findings from the technical expert panel, focus groups, and public comment suggest that the adaptation of this measure for the SMI population has specifications that can produce valid results. The high inter-rater reliability at the critical data element level also provides confidence in the validity of the measure.**

**In addition, the validity of the proposed measure is also enhanced by building on an existing measure, Preventive Care and Screening: Body Mass Index (BMI) Screening and Follow-Up (NQF # 0421), which is used in programs such as Physician Quality Reporting System, and CMS EHR Incentive program (“Meaningful Use”) and recommended for evaluating care for people with Dual Eligibility by the NQF Measure Applications Partnership Duals workgroup.**

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

**2b3. EXCLUSIONS ANALYSIS**

**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*)

**Claims and chart review data were used to identify exclusions. The claims codes of pregnancy are widely used across practitioners and settings and are considered to be valid. We tested the inter-rater reliability of the pregnancy exclusion through double abstraction of patient records.**

**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*)

**Twenty eight of 884 patients or 3.2% were excluded due to pregnancy. For the double abstracted sample, both abstractors identified the same individuals who met the exclusions. This yielded a Kappa statistic of 1 on measure exclusions. The performance scores with (37.6%) and without exclusion (37.9%) are nearly the same.**

**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 exclusion can be captured reliably in claims data without undue burden.**

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

**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

**2b4.2. If an outcome or resource use 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.**

**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**

**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)*

**Testing results (N = 3 health plans) did not provide sufficient data to conduct statistical tests.** **While the field test results are limited to 3 health plans, the findings suggest meaningful differences in performance are likely to exist across Medicaid plans.**

**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*)

**Table 7 shows the screening results and illustrates how the performance measure is calculated. Overall, 50.9% of patients were screened for obesity (range, 18% to 76%) and about half were identified as having a BMI ≥30 (49%, range 39% to 55%). Of these, less than half (39%) received any follow up care. The measure performance rate (35.6%) is calculated by adding patients who screened negative (25.9%) together with those who screened positive and received follow up care (9.7%). There was wide variation across the three plans, with final performance rates of 52.4%, 45.5%, and 10.9%. Within plans, the performance differences were similar across the different patient groups (SMI only, SMI + Diabetes, SMI + Hypertension).**

**Table 8 shows the performance rate by type of patient records. Health plan performance rates were similar when they were based on medical record results only (35.6%) versus results based on combined medical and behavioral health records (37.6%).**

**The performance rate of this measure as specified also varied by age and mental health diagnosis as shown in Table 9. The field test data suggested differences in performance between the two age groups (18-50 years versus >50 years). However, this result appeared to be related to variations in the age of patients across plans, since patients older than 50 years were mostly from the Dual SNP plan, which had the highest performance on this measure. Individuals with bipolar I disorder were less likely to receive BMI screening and follow-up compared to individuals with schizophrenia or major depression.**

**Table 7: Performance Rate (Medical Records Only)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **All Plans**  **(n=884)** | **Dual SNP**  **(n=345)** | **Medicaid Disabled**  **(n=219)** | **Medicaid Adult**  **(n=320)** |
| **Exclusions** | 28 | 13 | 6 | 9 |
| **Denominator after exclusions** | 856 | 333 | 213 | 311 |
| **Screening Results** |  |  |  |  |
| **Screened** | 436 (50.9%) | 253 (76.2%) | 127 (59.6%) | 56 (18.0%) |
| **Screened positive (BMI ≥30)** | 214 (49.1%) | 139 (54.9%) | 49 (38.6%) | 26 (46.4%) |
| **Received Follow-up among those that screen positive** | 83 (38.8%) | 60 (43.2%) | 19 (38.8%) | 4 (15.4%) |
| **Performance Rate** |  |  |  |  |
| **Screened negative** | 25.9% | 34.3% | 36.6% | 9.6% |
| **Screened positive with follow-up** | 9.7% | 18.1% | 8.9% | 1.3% |
| **Overall performance rate (screened negative + screened positive with follow up)** | 35.6% | 52.4% | 45.5% | 10.9% |

**Table 8. Performance Rate Using Medical Records Only Versus Combined Medical and Behavioral Health Records**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **All Plans**  **Rate (%)** | **Dual SNP**  **Rate (%)** | **Medicaid Disabled**  **Rate (%)** | **Medicaid Adult**  **Rate (%)** |
| **Medical record only** | 35.6 | 52.4 | 45.5 | 10.9 |
| **Both Medical and behavioral health records** | 37.6 | 54.8 | 49.3 | 11.5 |

**Table 9. Performance by Age, Gender, and Diagnosis (Combined Medical and Behavioral Health Records)**

|  |  |
| --- | --- |
|  | **Performance Rate** |
| **Overall** | 37.6% |
| **Age** |  |
| **18 to 50 years (inclusive)** | 30.3% |
| **Greater than 50 years** | 48.3% |
| **Gender** |  |
| **Male** | 37.9% |
| **Female** | 37.4% |
| **Diagnosis** |  |
| **Schizophrenia** | 37.8% |
| **Bipolar I disorder** | 31.1% |
| **Major depression** | 43.7% |

**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?*)

**Our field test showed low overall performance on this measure (37.6%) and high variation across the three field test plans (ranging from 11.5 – 54.8%) for people with SMI.**

**Two sources allow for comparison to a broader population. The existing provider-level measure is used in the Physician Quality Reporting System. The average performance rate on the existing measure (NQF # 0421) for Accountable Care Organizations (ACOs) reporting under Medicare Shared Savings Program and Pioneer ACO Model in 2012 was 54.3%.**

**In addition, we compared the screening rate only (without follow up) to body mass index screening for the general population covered by plans that report to NCQA. The rate of body mass index screening was 67.5% for Medicaid plans, with the with 10th percentile at 48.7% and 90th percentile of 84.4% compared to the screening rate in the field test plans of 50.9% (ranging from 18.0 – 76.2%).**

**Stakeholders reported that these findings were likely to be representative. Thus, we interpret the results to suggest that meaningful differences in performance exist and that there is substantial opportunity for improvement.**

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

**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 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**

**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*)

**Using standard rules of health plan reporting, sampled patients whose medical records are not available for review are considered numerator failures.**

**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.**