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

**Measure Number** (*if previously endorsed*)**: 2605(New Measure)**

**Measure Title**: **Follow-up after Discharge from the Emergency Department for Mental Health or Alcohol or Other Drug Dependence**

**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 specifications** (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*).

**Medicaid claims; Medicaid Analytic eXtract (MAX)**

**1.3. What are the dates of the data used in testing**? **Calendar year 2008**

**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: state | ☒ other: state |

**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, VALIDITY AND MEANINGFUL DIFFERENCES**

**We tested the reliability, validity, and variation in performance on this measure among 16 states for the rate of follow-up for mental health (MH) emergency department visits and 15 states for the rate of follow-up for alcohol and other drug dependence (AOD) emergency department visits using fee-for-service (FFS) Medicaid claims derived from the MAX data. We used FFS claims because Medicaid managed care organizations do not submit encounters in many states or submit incomplete data that limits the ability to observe every medical or behavioral health encounter.**

**We excluded states where FFS data were not expected to be representative (e.g. where only a small percentage of Medicaid adults were enrolled in FFS), where there was a problem with the Medicaid enrollment file or with FFS claims (e.g. inability to identify our population of interest, or missing claims), or where the denominator size of emergency department discharges was very small (less than 150).**

**Systematic Evaluation of 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*)

**Our analysis includes all Medicaid enrollees ages 18 and over. We excluded enrollees for whom Medicaid data would not be expected to include all instances of care provision including individuals who were (1) dually eligible for Medicare, (2) did not have full Medicaid benefits, (3) had private insurance, or (4) were enrolled in Medicaid for less than one calendar year.**

**The measure is calculated for two populations: (1) patients with a mental health emergency department visit and (2) patients with an alcohol or other drug dependence emergency department visit. For each population, there are two rates – follow-up within 7 days of emergency department discharge and follow up within 30 days of emergency department discharge. Table 1 summarizes the number and characteristics of individuals used to calculate the rates.**

**Table 1. Characteristics of patients in each denominator across all states included in analysis:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Mental Health Denominator** | | **AOD Denominator** | |
| **Number of states** | **N = 16** | | **N = 15** | |
| **Characteristic** | **Number** | **Percentage** | **Number** | **Percentage** |
| **Total Individuals** | **26,982** | **100** | **11,743** | **100** |
| **Gender** |  |  |  |  |
| **Male** | **10,744** | **39.8** | **6,068** | **51.7** |
| **Female** | **16,238** | **60.2** | **5,675** | **48.3** |
| **Unknown** | **0** | **0.0** | **0** | **0.0** |
| **Age** |  |  |  |  |
| **15 to 20** | **2,015** | **7.5** | **550** | **4.7** |
| **21 to 44** | **15,602** | **57.8** | **5,447** | **46.4** |
| **45 to 64** | **9,214** | **34.1** | **5,656** | **48.2** |
| **65 to 74** | **132** | **0.5** | **84** | **0.7** |
| **75 to 84** | **17** | **0.1** | **6** | **0.1** |
| **85+** | **2** | **0.0** | **0** | **0.0** |
| **Race/Ethnicity** |  |  |  |  |
| **African American** | **8,920** | **33.1** | **3,324** | **28.3** |
| **Caucasian** | **15,144** | **56.1** | **6,934** | **59.0** |
| **Hispanic** | **883** | **3.3** | **326** | **2.8** |
| **Other** | **485** | **1.8** | **377** | **3.2** |
| **Unknown** | **1,550** | **5.7** | **782** | **6.7** |
| **Medicaid Eligibility category** |  |  |  |  |
| **Adult** | **3,877** | **14.4** | **1,876** | **16.0** |
| **Disabled** | **22,439** | **83.2** | **9,575** | **81.5** |
| **Children** | **666** | **2.5** | **292** | **2.5** |
| **Geography** |  |  |  |  |
| **Metropolitan** | **11,146** | **41.3** | **5,021** | **42.8** |
| **Micropolitan** | **7,887** | **29.2** | **3,315** | **28.2** |
| **Neither** | **7,845** | **29.1** | **3,383** | **28.8** |
| **Unknown** | **104** | **0.4** | **24** | **0.2** |

**Source: MAX data from calendar year 2008**

**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 number of states used for each denominator is different; 16 states were included in our analysis of the follow-up rate for emergency department visits for mental health diagnoses whereas 15 states were included in our analysis of the follow-up rate for emergency department visits for AOD diagnoses. As seen in Table 2, The District of Columbia was not included in the AOD analysis due to a small sample size. There were no other differences in the data used for each aspect of testing.**

**Table 2: Number of emergency department discharges included in each denominator, by state:**

|  |  |  |
| --- | --- | --- |
| **State** | **Number of ED discharges in Mental Health Denominator** | **Number of ED discharges in AOD Denominator** |
| **AK** | **221** | **212** |
| **AL** | **2,294** | **873** |
| **CT** | **1,608** | **1,135** |
| **DC\*** | **181** | **N/A** |
| **GA** | **3,506** | **1,273** |
| **IL** | **5,681** | **1,248** |
| **IN** | **990** | **563** |
| **KY** | **3,520** | **1,403** |
| **LA** | **2,447** | **1,081** |
| **MN** | **2,149** | **747** |
| **MS** | **842** | **392** |
| **NC** | **4,907** | **2,416** |
| **NH** | **574** | **188** |
| **OK** | **813** | **514** |
| **WI** | **1,041** | **588** |
| **WV** | **1,178** | **704** |
| **Total** | **31,952** | **13,337** |

**\*DC was dropped from AOD denominator due to small sample size.**

**Source: MAX calendar year 2008**

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

**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 Testing of Performance Measure Score: In order to assess measure precision in the context of the observed variability across accountable entities, we used the beta-binomial method and resulting estimate described by Adams (2009). The following is quoted from the tutorial: “Reliability describes how well one can confidently distinguish the performance of one physician [or accountable entity] from another. Conceptually, it 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.” This approach is also relevant to health plans, states, and other accountable entities.**

**Adams’ approach uses a beta-binomial model to estimate reliability; this model is suited for estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach assumes that the performance measure score (pass/fail rate) across accountable entities has a flexible beta distribution, characterized by a signal variance. Given its performance measure score, the observed data (number of passes/failures) for an accountable entity has a binomial distribution, which provides the noise (measurement error) variance. From the beta-binomial model, the signal and noise variances are used to calculate reliability as:**

**Signal variance / (signal + noise variance)**

**Reliability scores vary from 0.0 to 1.0. A score of zero indicates that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 indicates that all variation is attributable to real differences in performance across accountable entities.**

**Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009**

**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*)

**Reliability statistic for follow-up for MH emergency department visits:**

**Average, 7-day follow-up: .99**

**10th-90th percentile across states: .97 – 1.0**

**Average, 30-day follow-up: .98**

**10th-90th percentile across states: .95 – 1.0**

**Reliability statistic for follow-up for AOD emergency department visits:**

**Average, 7-day follow-up: .99**

**10th-90th percentile across states: .98 – 1.0**

**Average, 30-day follow-u: .99**

**10th-90th percentile across states: .98 – 1.0**

**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?*)

**Reliability Testing of Performance Measure Score: Reliability scores can vary from 0.0 to 1.0. Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The testing suggests the all four follow-up rates reported as part of this measure have strong reliability between .98 and .99.**

**The minimum state-level reliability scores for this measure all exceed the minimally accepted threshold of 0.7.**

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

**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)*

**Empirical validity testing**

**We tested for construct validity by exploring whether states’ performance on this measure was related to their rates of inpatient hospitalization for mental health diagnoses (for the mental health denominator) or for alcohol and other drug use disorders (for the AOD denominator).  We hypothesized that states’ with lower rates of follow-up after discharge from the emergency department might have higher rates of inpatient stays for mental health and AOD. To evaluate the relationship between state performance on our measure and the state-level rate of inpatient stays, we fit a mixed effects logistic regression model. We regressed a beneficiary-level indicator of inpatient stay on a state-level binary variable indicating lowest vs. highest quartile performance follow-up after emergency department measure. To this we added a random effect of state to account for clustering of patients within states. If the p-value for the performance indicator variable is less than 0.05, then there is a significant difference in the rates of inpatient stays between states in the lowest vs. highest quartile of performance. If the p-value is greater than 0.05, then there is not a significant difference between low- and high-performing states.**

**Systematic Assessment of Face Validity**

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

**A multi-stakeholder technical expert panel of 16 individuals consisting of health plan representatives, behavioral health 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 multi-stakeholder focus groups that included 29 representatives from Medicaid plans, states, 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*)

**Table 3: Utilization of Inpatient Hospitalization for Mental Health Diagnosis by Measure Performance Quartile**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Enrollees Hospitalized for Mental Health Diagnosis (Percentage)** | |  |
|  | **Among States in Bottom 25 Percent of performance on FUED - Mental Health Denominator** | **Among States in Top 25 Percent of FUED – Mental Health Denominator** | **p-value** |
| **7-day follow-up** | **1.95%** | **1.58%** | **0.50** |
| **30-day follow-up** | **1.87%** | **1.64%** | **0.68** |

**Table 4: Utilization of Inpatient Hospitalization for AOD Diagnosis by Measure Performance Quartile**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Enrollees Hospitalized for AOD Diagnosis (Percentage)** | |  |
|  | **Among States in Bottom 25 Percent of FUED – AOD Denominator** | **Among States in Top 25 Percent of FUED - AOD Denominator** | **p-value** |
| **7-day follow-up** | **0.32%** | **0.35%** | **0.65** |
| **30-day follow-up** | **0.32%** | **0.33%** | **0.84** |

**Systematic assessment of face validity**

**Focus group stakeholders and the technical expert panel both supported the face validity of the measure. Both groups agreed that the transition period post-emergency room discharge was a critical time to get patients into outpatient care. Of the stakeholders who provided public comment for this measure, 18 total comments were received and 13 (72.2%) supported or supported the measure with modifications.** **Other commenters who did not support the measure had concerns about identifying whether an emergency visit took place as well as the validity of the emergency department diagnosis. Specifically, stakeholders were concerned that if the diagnosis in formation is not received, follow-up There were additional concerns about the ability to act on the 7-day follow-up as there is lag time between the date of the visit and when the claim is received by the organization. However, our multi-stakeholder expert panel recommended moving forward with the measure because the specifications and testing results were reasonable and the measure addresses important quality opportunity.**

**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?*)

**While the empirical testing did not support our hypothesis, stakeholders generally supported the face validity of the measure. The rate of inpatient hospitalization is not statistically different between states that perform well on this measure versus states that perform poorly (Tables 3 and 4). However, this result is likely due to the relatively low, tightly distributed rates of inpatient hospitalization for states in both the low- and high-performing groups. Rather than suggest that the measure is not valid, this result may indicate that our assumptions were not correct about the relationship between the measure and inpatient hospitalization; this relationship may warrant further study. The findings from public comment, focus groups and technical expert panel suggest that the adaptation for monitoring follow up after ED visits has specifications that can produce valid results.**

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

**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*)

**Our testing addresses four components of the denominator or exclusions, as shown in Table 5.**

**Table 5: Measure Exclusions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Exclusion** | **Rationale** | **MH Denominator lost due to exclusion** | **AOD Denominator lost due to exclusion** |
| 1. **ED discharges after December 1** | **If an ED discharge is after December 1, then the full 30-day follow-up period is not available for patient to receive follow-up care during the measurement year** | **7.5%** | **6.9%** |
| 1. **ED discharges who die during the follow-up period** | **Death prevents follow-up care** | **Less than 1%** | **Less than 1%** |
| 1. **For an ED discharge where the patient also visited the ED in the previous 30 days, exclude those previous ED discharges** | 1. **Including these ED discharges could lead to a larger number of ED visits resulting in higher performance on the measure** 2. **This exclusion aligns with the NQF-endorsed (#0576) Follow-up after Hospitalization for Mental Illness measure to reduce the burden and confusion for health plans implementing both measures** | **16.2%** | **17.3%** |
| 1. **ED discharges with an inpatient or other residential stay during follow-up period** | 1. **An inpatient or otherwise residential stay may interfere with the receipt of outpatient follow-up care** 2. **This exclusion aligns with the NQF-endorsed (#0576) Follow-up after Hospitalization for Mental Illness measure to reduce the burden and confusion for health plans implementing both measures** | **34.2%** | **40.8%** |

**Note: The exclusions presented in this table are not mutually exclusive. For example, a discharge that falls under exclusions 1 and 4 would appear in both places in this table.**

**We tested whether the exclusions affected over performance scores.**

**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*)

**Table 6: Number and percent of denominator remaining after exclusions, by state**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mental Health (MH) Denominator** | | | **AOD Denominator** | | |
| **State** | **MH denominator before exclusions** | **MH denominator after exclusions** | **Percent after exclusions** | **AOD denominator before exclusions** | **AOD denominator after exclusions** | **Percent after exclusions** |
| **AK** | **297** | **221** | **74.4%** | **294** | **212** | **72.1%** |
| **AL** | **3,244** | **2,294** | **70.7%** | **1,135** | **873** | **76.9%** |
| **CT** | **2,800** | **1,608** | **57.4%** | **2,081** | **1,135** | **54.5%** |
| **DC\*** | **311** | **181** | **58.2%** | **302** | **0** | **0.0%** |
| **GA** | **5,009** | **3,506** | **70.0%** | **1,796** | **1,273** | **70.9%** |
| **IL** | **11,057** | **5,681** | **51.4%** | **3,179** | **1,248** | **39.3%** |
| **IN** | **1,405** | **990** | **70.5%** | **765** | **563** | **73.6%** |
| **KY** | **4,762** | **3,520** | **73.9%** | **1,879** | **1,403** | **74.7%** |
| **LA** | **3,738** | **2,447** | **65.5%** | **1,451** | **1,081** | **74.5%** |
| **MN** | **3,192** | **2,149** | **67.3%** | **1,100** | **747** | **67.9%** |
| **MS** | **1,198** | **842** | **70.3%** | **524** | **392** | **74.8%** |
| **NC** | **6,755** | **4,907** | **72.6%** | **3,372** | **2,416** | **71.6%** |
| **NH** | **800** | **574** | **71.8%** | **292** | **188** | **64.4%** |
| **OK** | **1,183** | **813** | **68.7%** | **717** | **514** | **71.7%** |
| **WI** | **1,491** | **1,041** | **69.8%** | **895** | **588** | **65.7%** |
| **WV** | **1,699** | **1,178** | **69.3%** | **934** | **704** | **75.4%** |
| **Total** | **48,941** | **31,952** | **65.3%** | **20,716** | **13,337** | **64.4%** |

**\*DC was dropped from AOD denominator due to small sample size.**

**Table 7: Measure performance before and after application of final exclusion**

|  |  |  |
| --- | --- | --- |
| **Measure** | **Overall measure performance after exclusions 1-3 applied** | **Overall measure performance after exclusions 1-4 applied** |
| **Mental Health: 7-day follow-up** | **64.6** | **66.0** |
| **Mental Health: 30-day follow-up** | **75.6** | **76.1** |
| **AOD: 7-day follow-up** | **61.9** | **64.3** |
| **AOD: 30-day follow up** | **65.9** | **66.7** |

**Note: The overall performance rates presented here are pooled across states.**

**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*)

**We tested several exclusions in order to understand the impact on the denominator. Exclusions 1 and 2 are necessary to ensure that follow-up care can be observed during the measurement year. Exclusion 3 is prevents incentivizing more emergency department visits and aligns with other NQF endorsed measures to decrease burden and confusion for health plans. Average measure performance does not change substantially when Exclusion 4 is implemented, this exclusion aligns with NQF measure 0576, and there is a clinical rationale for excluding emergency department discharges that have an inpatient or other residential stay during the follow-up period, which is important to the face validity of the measure. All of the exclusions have minimal effect on the burden of calculating the measure since these exclusions are derived exclusively from claims data. In the specifications, some of these exclusions have been incorporated into the denominator definition.**

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

**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)*

**Empirical testing**

**To demonstrate meaningful differences in performance, we calculated an inter-quartile range (IQR) for each rate. 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, we calculate a Chi-squared test of the performance difference between each state in the lowest quartile vs. each state in the highest quartile. The Chi-squared test method calculates a test statistic based on the sample size and performance rate of each state. If the p value of the test statistic is less than .05, then the two states’ performance is significantly different from each other. Using this method, we compared the performance rates of each pair of states, one state in the 25th percentile and another state in the 75th percentile of performance.**

**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 8: Variation in performance across states**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **10th** | **25th** | **Median** | **75th** | **90th** | **IQR** | **Minimum** | **Maximum** | **p-value** |
| **Mental Health: 7-day follow-up** | **42.2** | **59.5** | **73.8** | **79.5** | **90.0** | **20.0** | **35.5** | **89.5** | **<.001** |
| **Mental Health: 30-day follow-up** | **59.9** | **75.3** | **81.8** | **84.8** | **86.0** | **9.5** | **53.8** | **92.4** | **<.001** |
| **AOD: 7-day follow-up** | **21.8** | **49.5** | **68.5** | **80.3** | **83.1** | **30.8** | **15.5** | **90.3** | **<.001** |
| **AOD: 30-day follow up** | **28.7** | **52.2** | **70.6** | **80.8** | **83.9** | **28.5** | **26.8** | **90.3** | **<.001** |

**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?*)

**The results above indicate there is a gap in performance between the 25th and 75th performing states, ranging from 9.5 percentage points on the 7-day mental health measure to 30.8 on the 7-day AOD measure. For all states and all rates, the difference between the 25th and 75th percentile is statistically significant.**

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

**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*)

**This measure is collected using all available administrative claims; there are no missing data on this measure.**

**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.**