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

**Measure Number** (*if previously endorsed*)**:** 1879

**Measure Title**: Adherence to Antipsychotic Medications for Individuals with Schizophrenia

**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: | other: |

**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*).

Medicare Parts A, B, and D claims data and Minimum Data Set (MDS) data were used to support the field testing of the measure. The following files were used:

* Denominator tables to determine individual enrollment
* Prescription drug benefit (Part D) coverage tables
* Beneficiary file
* Institutional claims (Part A)
* Non-institutional claims (Part B)—physician carrier/non-DME (durable medical equipment)
* Prescription drug benefit (Part D) claims
* Centers for Medicare & Medicaid Services (CMS) physician and physician specialty tables
* National Plan & Provider Enumeration System (NPPES) database

**1.3. What are the dates of the data used in testing**? 2007, 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.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: population (state) | other: population (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*)

Data from eight states were included in the testing and analysis for validity and physician group and state reliability. These data included 9,406 Physician Groups and 656 Part D plans.

For health plan reliability testing, data included five randomly selected Part D plans from two states.

**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 data included 4,789,034 Medicare beneficiaries.

**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**.

No differences in the data or sample used.

**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.

Two proxy variables for social risk were evaluated to understand disparities: race/ethnicity and dual-eligibility beneficiary status. Because this measure is not an outcome or intermediate outcome measure, these factors were not evaluated for risk adjustment. Overall, in the younger age groups (18-64), African-Americans had noticeably lower adherence. In all age groups, dual-eligible beneficiaries had higher rates of adherence than those who are not dual-eligible.

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

**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*)

In order to assess measure precision in the context of the observed variability across measurement units (physician groups), we utilized the approach proposed by Adams (2009) in work on the reliability for provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician 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. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.”

The signal to noise ratio was calculated as a function of the variance between physician groups (signal) and the variance within a physician group (noise). Reliability was estimated using a beta-binomial model. This approach has 2 basic assumptions:

1) Each physician has a true pass rate, p, which varies from physician to physician, and

2) The physician’s score is a binomial random variable conditional on the physician’s true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual physician group variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across physician groups). In a simulation, Adams showed that differences between physicians started to be seen at reliability of 0.7 and significant differences could be seen at reliability of 0.9. Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between physicians. Reliability scores were also calculated for state level results using the same approach.

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

Reliability at the health plan level was assessed using Cohen’s Kappa. The measure scores for five randomly selected Medicare Part D plans were compared and inter-rater agreement was calculated. Concerning an acceptable threshold for kappa, there are no definitive criteria in the literature for what level of reliability is acceptable for measures based on administrative data. Furthermore, since relatively small differences in programmer interpretation could result in a large variation in output, we utilized a conservative threshold of 0.9 for Cohen’s Kappa, based on the following scale:

< 0 = no agreement

0–0.20 = slight agreement

0.21–0.40 = fair agreement

0.41–0.60 = moderate agreement

0.61–0.80 = substantial agreement

0.81–1 = almost perfect agreement

Therefore, if the Cohen’s Kappa was greater than or equal to 0.9, the measure specifications were considered reliable. If Cohen’s Kappa in the initial reliability testing with the two programmers was less than 0.9, each step of the measure algorithm (in the Measure Information Form [MIF]) was compared, and the differences were clarified between programmer 1 and 2. Identified differences are noted in a narrative, where applicable, along with extracts of the respective modification to the MIF.

The revised MIF was then presented to a third programmer and results compared to the consolidated results derived in the first round of reliability testing. This iterative process with independent programmers continued until the Kappa score reached the threshold of greater than or equal to 0.9.

**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*)

**State Reliability**

State / Denominator / Mean rate for state / Reliability score (based on the mean rate)

A / 1368 / 67.54% / 0.955

B / 681 / 76.36% / 0.927

C / 14869 / 71.03% / 0.996

D / 3652 / 84.72% / 0.990

E / 6157 / 80.02% / 0.992

F / 3351 / 68.49% / 0.981

G / 1005 / 78.31% / 0.952

H / 5224 / 81.13% / 0.991

**Physician Group Reliability (By Case Volume)**

Minimum denominator size of MD group / # of Groups / Mean rate of physician groups / Variance between physician groups / Physician specific error / Reliability score (based on the mean rate and the minimum denominator size) / Mean Reliability Score / Median Reliability score / Minimum Reliability Score / Maximum Reliability Score / Standard Deviation of Reliability Scores

10 / 296 / 76.71% / 0.0081 / 0.0179 / 0.3116 / 0.48 / 0.44 / 0.26 / 0.91 / 0.15

20 / 122 / 77.49% / 0.0087 / 0.0087 / 0.4993 / 0.65 / 0.63 / 0.43 / 0.91 / 0.13

30 / 71 / 79.08% / 0.0079 / 0.0055 / 0.5895 / 0.71 / 0.71 / 0.52 / 0.91 / 0.11

35 / 55 / 80.28% / 0.0081 / 0.0045 / 0.6405 / 0.75 / 0.74 / 0.58 / 0.91 / 0.09

40 / 44 / 80.94% / 0.0088 / 0.0039 / 0.6954 / 0.79 / 0.8 / 0.64 / 0.92 / 0.08

45 / 36 / 81.41% / 0.0084 / 0.0034 / 0.7144 / 0.8 / 0.8 / 0.67 / 0.92 / 0.07

50 / 30 / 80.68% / 0.0092 / 0.0031 / 0.7471 / 0.82 / 0.83 / 0.69 / 0.92 / 0.06

100 / 7 / 74.55% / 0.0194 / 0.0019 / 0.9107 / 0.94 / 0.95 / 0.91 / 0.96 / 0.02

150 / 3 / 75.47% / 0.0032 / 0.0012 / 0.7186 / 0.75 / 0.76 / 0.71 / 0.77 / 0.03

**Health Plan Reliability**

|  | **Percent Agreement** | |  |
| --- | --- | --- | --- |
| **Unit of Analysis** | **Programmer 1**  **Num/Den (%)** | **Programmer 2**  **Num/Den (%)** | **Final Cohen’s Kappa** |
| **Part D Plan 1** | 44/75 (58.7%) | 45/75 (60.0%) | 0.97 |
| **Part D Plan 2** | 478/677 (70.6%) | 459/675 (68.0%) | 0.93 |
| **Part D Plan 3** | 74/109 (67.9%) | 72/109 (66.1%) | 0.96 |
| **Part D Plan 4** | 49/71 (69.0%) | 48/71 (67.6%) | 0.97 |
| **Part D Plan 5** | 49/63 (77.8%) | 48/63 (76.2%) | 0.95 |

**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?*)

**State Reliability**

All state-level reliability scores were > 0.9; indicating that the measure would produce reliable scores at the state level.

**Physician Group Reliability**

The original denominator threshold tested was 30 patients, resulting in 53.5% (N=38) of 71 physician groups attributed having reliable scores (defined as 0.7 or greater). Increasing the denominator size to 45 patients resulted in 94.4% (N=34) of 36 physician groups with a reliable score. Among these groups, overall reliability was 0.71, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between physician groups. Therefore, these results suggest that physician groups with 45 patients or more will produce reliable scores.

**Health Plan Reliability**

Results obtained by the final two independent programmers met the Kappa threshold of 0.9, and no further refinement of measure specification was deemed necessary.

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

**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.

Empirical validity testing is not available for this measure at the time of this maintenance review. Analysis was not possible in the timeframe from NQF publication of this new evaluation criteria (September 2017) and submission of the testing form (January 2018). On March 9, 2018, the measure steward, CMS, met with NQF to discuss submission of this measure. NQF requires empiric validity testing at the time of maintenance; however, they recognize the limitations of the timeframe for submission. NQF, CMS, and the contract team agreed that in leu of providing results of testing, it would be suitable to include a detailed plan for testing empiric validity before the next maintenance submission.

We will test measure performance score validity by examining correlations with meaningful measures of a similar quality construct (convergent validity) using the Spearman’s rank correlation coefficient. We will analyze the convergent validity of the measures, evaluating the extent to which the measures *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) and *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) correlate. We hypothesize that health plans and provider groups that perform well at helping individuals with schizophrenia remain adherent to antipsychotic medications will also perform well at helping individuals with bipolar I disorder remain adherent to mood stabilizers. Both measures are indicators of overall quality of care for individuals with serious mental illness and should be correlated.

For health plan level testing, we will evaluate the correlation between *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) and *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) using Medicare-Medicaid Plan (MMP) encounter data. We will begin our initial testing using data already available to us from federal fiscal years 2015 and 2016, covering dates between October 1, 2014, and September 30, 2016. Because of the uncertain quality of the encounter data reported by MMPs, we will conduct an initial series of data checks to examine the quality and volume of encounter data required for the measures and include MMPs for which the quality is sufficient for testing purposes. Our initial data checks will examine quality and volume of data at the plan and state levels to ensure sufficient sample sizes for testing the research questions. We anticipate using data elements related to Medicare and Medicaid enrollment, institutional encounters, non-institutional encounters, and prescription drug coverage and claims.

For provider level testing we will evaluate the correlation between *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) and *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) using Medicare FFS data paired with Medicare Part D claims data. We will pull this Medicare FFS data from the Integrated Data Repository (IDR) to complete testing. No Medicaid data will be used. We anticipate using data elements related to Medicare and Medicaid enrollment, institutional claims, non-institutional claims, and prescription drug coverage and claims.

We will produce scatter plots comparing the two measures at the provider and health plan level. The Spearman’s rank correlation coefficient (rs) assesses the monotonic relationship in plan rankings for each measure pair. The coefficient ranges from -1 to 1, where rs = 1 indicates perfect alignment of plan rankings, rs = – 1 indicates opposite alignment of plan rankings, and rs = 0 represents no alignment in plan rankings. We will fit a smooth curve using locally weighted scatterplot smoothing (LOWESS) method to visualize any trends in the scatterplots. Because the LOWESS method does not rely on a preconceived model for the distribution of the measures (non-parametric), the LOWESS curve can captured detailed information about the measure relationships that the correlation coefficient does not convey.

The timeline for this work is described below:

* October – November 2018: Develop analytic file
* November 2018 – February 2019: Conduct validity testing and review results
* March – April 2019: Summarize results and update measure documentation
* TBD: Submit updated validity testing to NQF as part of maintenance submission

Although empiric validity analysis has not yet been conducted, this measure uses a definition of adherence (0.8 proportion of days covered) that is harmonized with other National Quality Forum (NQF) endorsed adherence measures and is consistent with the threshold of adherence used in the seven studies cited in the evidence attachment. These studies demonstrated improved outcomes in schizophrenia associated with adherence to medication. Although many of these studies have used the medication possession ratio (MPR) rather than the proportion of days covered (PDC), CMS and the Pharmacy Quality Alliance (PQA) have evaluated and extensively tested the PDC and the MPR and specifically found that: 1) the PDC and MPR will provide nearly identical results when examining adherence to a single drug; 2) the PDC will provide a more conservative estimate of adherence when examining adherence to a class of drugs that are prone to frequent switching and concomitant therapy with multiple drugs within the class (as with antipsychotic drugs). Therefore, based on NQF’s recommendation that a standard methodology for calculating medication adherence be established across all endorsed adherence measures, CMS and PQA agreed to harmonize the methodology for calculating medication adherence using the PDC, which was approved by the NQF Consensus Standards Approval Committee (CSAC).

**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)*

**Face Validity**

A Technical Expert Panel (TEP), comprising internal medicine physicians and pharmacists, evaluated the face validity of the measure and measure scores. The following 12 TEP members evaluated the face validity of the measure and measure scores:

1. Jill S. Borchert, Pharm.D., BCPS, Professor, Pharmacy Practice & PGY1 Residency Program Director, Midwestern University, Chicago College of Pharmacy

2. Anne Burns, RPh, Vice President, Professional Affairs, American Pharmacists Association

3. Jannet Carmichael, Pharm.D., FCCP, FAPhA, BCPS, VISN 21 Pharmacy Executive, VA Sierra Pacific Network

4. Marshall H. Chin, MD, MPH, Professor of Medicine, University of Chicago

5. Jay A. Gold, MD, JD, MPH, Senior Vice President and Medicare Chief Medical Officer, MetaStar, Inc.

6. David Nau, Ph.D., R.Ph., CPHQ, Senior Director of Research & Performance Measurement, PQA, Inc.

7. N. Lee Rucker, M.S.P.H., Senior Strategic Policy Advisor, AARP - Public Policy Institute

8. Marissa Schlaifer, MS, RPh, Director of Pharmacy Affairs Academy of Managed Care Pharmacy

9. Brad Tice, Pharm.D., Chief Clinical Officer, PharmMD Solutions, LLC

10. Jennifer K. Thomas, Pharm.D., Manager, Pharmacy Services, Delmarva Foundation for Medical Care/Delmarva Foundation of the District of Columbia

11. Darren Triller, Pharm.D., Director, Pharmacy Services, IPRO

12. Neil Wenger, MD, Professor of Medicine, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research

The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using Survey Monkey). Face validity of the measure score as an indicator of quality was systematically assessed as follows: After the measure was fully specified and tested, the expert panel members were asked to rate, based on a 5-point Likert scale, their level of agreement with the following statement: "The measure appears to measure what is intended."

The 5-point Likert scale was defined as follows: 1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree

**ICD-10-CM Conversion Methodology**

The conversion of the measure to include ICD-10-CM codes is provided as requested by NQF. The crosswalk is provided as an excel file in Section S2.b Data Dictionary or Code Table.

Name and Credentials of Experts Who Assisted in the Process

* Soeren Mattke, MD, DSc, Senior Scientist, RAND Corporation
* Tim Laios, MBA, MPH, Executive Director, Informatics, Health Services Advisory Group (HSAG)
* Ryan Fair, BS, Director, Informatics, HSAG
* Kerri Carlile, MS, Informatics Analyst, HSAG
* Sara Lomeli, BA, Informatics Project Coordinator, HSAG

Evaluation of ICD-9-CM Changes

The changes (i.e., deletions and/or additions) made to the ICD-9-CM codes for the measures requiring conversion were reviewed. Additionally, the ICD-9-CM codes were reviewed for any coding updates, using the October 2011 Conversion Table of New ICD-9-CM Codes, published by the National Center for Health Statistics (NCHS) and the Centers for Medicare & Medicaid Services (CMS).

ICD-9-CM Code Identification

For each measure requiring conversion, original tables were used to identify all ICD-9-CM codes included in the measure. Those ICD-9-CM codes and matching descriptions were then extracted from the Ingenix 2011 ICD-9-CM Data File. Only valid ICD-9-CM codes were retained and used in the ICD-9-CM to ICD-10-CM conversion process.

Ingenix Extraction

When extracting the ICD-9-CM codes from the Ingenix Data File, all codes were extracted with two-decimal specificity. For example, for ICD-9-CM code 274.1, all ICD-9-CM codes that had 2741 for the first four digits were extracted (e.g., 274.10, 274.11, and 274.19). For every three-digit ICD-9-CM code used in the measure, all ICD-9-CM codes that began with those first three digits were extracted. For the ICD-9-CM codes listed in ranges, codes with up to two-decimal specificity were extracted within that range.

Conversion Process

The ICD-9-CM and ICD-10-CM General Equivalence Map (GEM) text files and the ICD-10-CM Descriptions text file were imported into SAS and combined into one data file to capture all ICD-9-CM codes, their corresponding ICD-10-CM codes, and the ICD-10-CM code descriptions. The ICD-9-CM codes that were retained from the Ingenix 2011 ICD-9-CM Data File described above were then extracted from the combined GEM data file.

The results for each measure were then exported into Excel and validated to ensure that the translation was appropriate (i.e., the crosswalk was correct and applied appropriately and all appropriate ICD-9-CM codes were captured). Since one ICD-9-CM code can have several corresponding ICD-10-CM codes, each ICD-9-CM code can have multiple entries in the final Excel document (i.e., one row for each corresponding ICD-10-CM code).

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

**Systematic Assessment of Face Validity**

The results of the Technical Expert Panel rating of face validity as represented by this statement, "The measure appears to measure what is intended," on a scale of 1 to 5.

N=12 panel members, Mean Rating=4.33

Response / % of TEP / Number of TEP

Strongly Agree / 33.3% / 4

Agree / 66.7% / 8

Neutral / 0.0% / 0

Disagree / 0.0% / 0

Strongly Disagree / 0.0% / 0

**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?*)

In summary, 100% of the TEP members responded “agree” or “strongly agree” with the statement that the measure, as specified, had face validity. The results indicate strong support of the face validity of the measure by the Technical Expert Panel.

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

**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*)

**Type of analysis**

A sensitivity analysis was conducted to estimate the effect of the exclusion on the overall measure rate across the eight-state sample. The overall prevalence of the exclusion was calculated and the measure rate was calculated two ways: 1) with the exclusion applied and 2) without the exclusion applied.

**Description of exclusion**

Individuals with a diagnosis of dementia were excluded from the measure denominator. In April 2005, the Food and Drug Administration (FDA) issued a Public Health Advisory, which warned of the increased risk of mortality associated with the use of atypical antipsychotics in elderly patients with dementia. This warning was based on the findings of a meta-analysis of 17 short-term, randomized, placebo-controlled trials and showed that the risk of death in drug-treated patients was 1.6 to 1.7 times the risk of death in placebo-treated patients (Schneider et al., 2005). In 2008, the FDA Advisory and Black Box Warning was extended to all antipsychotic medications when further studies (Liperoti et al., 2009; Schneeweiss et al., 2007; Setoguchi et al., 2008) showed that conventional antipsychotics were associated with a similar increased risk of death when administered to elderly patients with a diagnosis of dementia.

Liperoti, R., Onder, G., Landi, F., Lapane, K. L., Mor, V., Bernabei, R., & Gambassi, G. (2009). All-cause mortality associated with atypical and conventional antipsychotics among nursing home residents with dementia: A retrospective cohort study. Journal of Clinical Psychiatry, 70(10),1340-1347.

Schneeweiss, S., Setoguchi, S., Brookhart, A., Dormuth, C., & Wang, P. S. (2007). Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ, 176, 627–632. [PubMed: 17325327]

Schneider, L. S., Dagerman, K. S., & Insel, P. (2005). Risk of death with atypical antipsychotic drug treatment for dementia: Meta-analysis of randomized placebo-controlled trials. Journal of the American Medical Association, 294, 1934–1943. [PubMed: 16234500]

Setoguchi, S., Wang, P. S., Brookhart, M., Canning, C. F., Kaci, L., & Schneeweiss, S. (2008). Potential causes of higher mortality in elderly users of conventional and atypical antipsychotic medications. JAGS, 56, 1644–1650.

**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*)

Individuals with dementia represented approximately 11% of all individuals in the measure denominator. If individuals with dementia were excluded, the measure rate was 74.4% (31,752/42,676) across the eight-state sample; whereas, the measure rate without excluding these individuals was 74.0% (35,416/47,852).

**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*)

While overall performance across the eight-state sample did not differ, individuals with dementia represent a population where adherence to antipsychotic medications is associated with an increased risk of mortality. Therefore, the Technical Expert Panel recommended excluding this subpopulation.

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

**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)*

To identify statistically significant differences in performance for states and physician groups, we conducted a comparison of means and percentiles. Confidence intervals (95%) were calculated around point estimates and then compared to the grand mean of states. If the confidence intervals did not overlap with the overall grand mean, the comparison was considered statistically significant.

For physician groups and health plans, the observed sample sizes of members of each comparison unit were tested empirically to determine whether there was sufficient power to detect statistically significant differences between members (e.g., between plans or between physician groups). To do this, all members were divided into quintiles according to their measure score. Within each quintile, the member with a denominator closest in size to the median denominator of the quintile and the member with the measure score closest to the median measure score of that quintile were identified. Comparison of members based on their median denominator size was made, because a relationship between denominator size and quality cannot be excluded a priori. In addition, a “standardized” member of each quintile was simulated by using the median denominator size across all quintiles. Binomial (exact) 95% confidence intervals for each of the 10 selected plans or physician groups (i.e., two plans or physician groups per quintile) were calculated around the point estimates. Overlapping confidence intervals indicate insufficient statistical power to detect statistically significant differences.

**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*)

**Meaningful Differences at the State Level**

Below we present the measure rate by state, mean, median, and standard deviation.

State A – 67.5%\* (statistically significantly lower than the mean)

State B – 76.4%

State C – 71.0%\* (statistically significantly lower than the mean)

State D - 84.7%\* (statistically significantly higher than the mean)

State E – 80.0%

State F - 68.5%\* (statistically significantly lower than the mean)

State G - 78.3%

State H - 81.1%

Mean of state scores – 75.9%

Median of state scores – 77.4%

Standard deviation of state scores – 6.3%

**Meaningful Differences at the Physician Group Level**

Below we present the mean, standard deviation, and percentiles at the physician group level.

Number of Physician Groups with at least 45 individuals in measure denominator = 36

Mean: 81.4%

SD: 10.8%

10th Percentile: 68.0%

25th Percentile: 77.9%

50th Percentile: 82.6%

75th Percentile: 89.0%

90th Percentile: 92.3%

Of physician group scores, 8.3% were statistically significantly lower than the mean, and 33.3% of physician group scores were statistically significantly higher than the mean, indicating a wide range of scores.

| **Across Physician Groups with ≥ 30 Beneficiaries** | **Quintile 1** | **Quintile 2** | **Quintile 3** | **Quintile 4** | **Quintile 5** |
| --- | --- | --- | --- | --- | --- |
| **Number of physician groups** | 6 | 5 | 6 | 6 | 5 |
| **Denominator range across physician groups**  **(minimum-maximum)** | 30-140 | 30-37 | 31-73 | 39-143 | 30-46 |
| **Median denominator size** | 50 | 34 | 37 | 55 | 42 |
| **Measure score (95% CI) of the physician group with a denominator size closest to the median denominator size** | 66.7%  (53.9-80.0) | 73.5%  (60.0-87.1) | 75.8%  (62.5-88.9) | 81.5%  (71.9-90.7) | 88.1%  (79.6-96.0) |
| **Measure score range across physician groups**  **(minimum-maximum)** | 37.9%-66.7% | 69.7%-73.5% | 75.0%-77.6% | 79.5%-83.6% | 85.7%-93.5% |
| **Median measure score** | 61.2% | 73.0% | 77.1% | 81.7% | 88.1% |
| **Measure score (95% CI) of the group with a score closest to the median score** | 63.1%  (54.2-72.4) | 73.0%  (59.9-86.2) | 77.4%  (64.2-90.4) | 81.5%  (71.9-90.7) | 88.1%  (79.6-96.0) |
| **95% CI using the overall median denominator N=42** | 61.2%  (47.5-75.8) | 73.0%  (60.6-85.5) | 77.1%  (65.4-88.6) | 81.7%  (71.0-91.9) | 88.1%  (79.6-96.0) |
| CI = Confidence Interval | | | | | |

**Meaningful Differences at the Health Plan Level**

| **Across Part D**  **Plan with ≥ 30 Beneficiaries** | **Quintile 1** | **Quintile 2** | **Quintile 3** | **Quintile 4** | **Quintile 5** |
| --- | --- | --- | --- | --- | --- |
| **Number of plans** | 4 | 5 | 5 | 5 | 4 |
| **Denominator range across plans (minimum-maximum)** | 34-220 | 97-1,267 | 238-3,188 | 792-3,304 | 53-413 |
| **Median denominator size** | 117 | 314 | 2,234 | 1,338 | 212 |
| **Measure score (95% CI) of the plan with a denominator size closest to the median denominator size** | 63.3%  (55.6-71.3) | 67.8%  (62.8-73.0) | 70.5%  (68.7-72.4) | 74.0%  (71.7-76.3) | 78.5%  (73.1-83.8) |
| **Measure score range across plans** | 58.8%-64.1% | 66.0%-67.8% | 69.4%-71.4% | 72.5%-75.6% | 77.4%-81.8% |
| **Median measure score** | 63.6% | 66.1% | 69.7% | 74.0% | 78.0% |
| **Measure score and 95% CI of the plan with a score closest to the median score** | 63.3%  (55.6-71.3) | 66.1%  (61.5-70.8) | 69.7%  (64.1-75.5) | 74.0%  (71.7-76.3) | 78.5%  (73.1-83.8) |
| **95% CI based on the overall median denominator size N=389** | 63.6%  (58.9-68.4) | 66.1%  (61.5-70.8) | 69.7%  (65.3-74.3) | 74.0%  (69.7-78.3) | 78.0%  (74.0-82.0) |
| CI = Confidence Interval | | | | | |

**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?*)

**Meaningful Differences at the State Level**

Three states (37.5%) had scores statistically significantly lower than the mean and one state (12.5%) had scores significantly higher than the mean. Measure rates by state ranged from 67.5% in state A to 84.7% in state D, indicating suboptimal performance across all states and variation between high- and low-performing states.

**Meaningful Differences at the Physician Group Level**

The testing results indicate ample room for improvement and meaningful differences in quality of care between the highest and lowest performing physician groups. Overall 41.6% of physician performance scores were statistically different from the mean. For those physician groups with at least 45 eligible individuals, high- (90th percentile) and low- (10th percentile) performing physician groups were 24.3 percentage points apart.

Please note after testing was conducted the measure was harmonized to include individuals receiving depot injections (rather than exclude those individuals). The testing data presented above do not yet reflect the change in specification. Our preliminary testing since the addition of individuals receiving depot injections showed that the impact of this inclusion increases the denominator size by approximately 23% and decreases the overall measure rate across the eight-state sample by 2.2 percentage points.

A total of 28 physician groups with at least 30 beneficiaries were identified and could be distributed across the measure score quintiles. Physician groups showed limited variation in sample size with no particular pattern with respect to measure scores. We noted pronounced variation in measure rates across physician groups, ranging from 37.9% to 93.5%, but denominator sizes were consistently small, resulting in wide confidence intervals. Comparison of standardized physician groups (calculated based on the score closest to the median measure score or the overall median denominator size) showed sufficient discriminatory ability between physician groups of the highest and lowest quintiles.

Assuming a median measure rate of 77.1% and a median denominator of 42 beneficiaries, the smallest statistically significant difference that can be detected at the physician group level with a power of 80% and α=0.05 is 18.0%.

**Meaningful Differences at the Health Plan Level**

A total of 23 plans with at least 30 beneficiaries could be distributed across the measure score quintiles. Plans showed pronounced variation in sample size with a general pattern in the first 4 quintiles of increasing size with respect to measure scores. Comparison of individual plans (selected based on the denominator size closest to the median denominator size or score closest to the median measure score) showed sufficient discriminatory ability, based on lack of overlap between the confidence intervals of the lowest and highest performing quintiles and limited discriminatory ability between the lowest quintile and the 4th quintile. Comparison of standardized plans (with confidence intervals calculated based on the overall median denominator size of the entire sample) showed sufficient discriminatory ability between members of the highest and lowest quintiles, as well as between the lowest quintile and the 4th quintile. Of note, the sample sizes for plans varied dramatically within each quintile and will result in distinctly different power if two members are compared.

Assuming a median measure rate of 69.7% and a median denominator of 389 beneficiaries, the smallest statistically significant difference in measure rates that can be detected at the plan level with a power of 80% and α=0.05 is 6.9%.

Please note after testing was conducted the measure was harmonized to include individuals receiving depot injections (rather than exclude those individuals). The testing data presented above do not yet reflect the change in specification. Our preliminary testing since the addition of individuals receiving depot injections showed that the impact of this inclusion increases the denominator size by approximately 23% and decreases the overall measure rate across the eight-state sample by 2.2 percentage points.

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

**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*)

Missing days’ supply data and bias from cash prescriptions were possible threats to validity. An empirical assessment of these possible threats was conducted as follows:

Threat of Bias from Missing Data

We have identified two potential scenarios where measure results could be biased by missing data:

1. Missing days’ supply within the prescription drug event data, which is a required data element to calculate medication adherence;
2. Missing drug claims due to individuals using alternative payment sources for prescription drugs, e.g., $4 commercial discount prescription programs and other alternative drug benefits, such as the Veterans Administration (VA)

For missing days’ supply, we analyzed the number (%) of beneficiaries in the measure denominator with one or more claims that had missing days’ supply.

For bias from cash prescriptions or alternative sources, we conducted a limited sensitivity analysis using a two-state sample (states C and G) to estimate the potential impact of a commercial cash discount program on measure rates. Specifically, we created a National Drug Code (NDC) list from the formulary of a leading cash discount program to identify those individuals with at least one claim for an antipsychotic on the formulary and no claims for any other Part D drugs on the formulary as a proxy to potentially indicate the individual was filling medications through the cash discount program. We then simulated the effect on measure rates, if each of these individuals’ antipsychotic drug use extended from the last consecutive claim to the end of the measurement period, assuming that individuals had switched to the cash program. We simulated two scenarios: including complete coverage of all remaining days’ until the end of the measurement period were 100% or extrapolating the average proportion of days covered from the first prescription in the measurement period to the last prescription in the measurement period.

**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*)

**Missing Data**

Days' Supply: Only 2 individuals (0.005%) in the overall measure denominator had one or more claims with missing days’ supply.

**Cash Prescriptions**

The percentage of individuals in the denominator with antipsychotic Part D claims on the formulary and no claims for any other drugs on the commercial discount formulary was 0.9% (145/15,874).

SCENARIO 1. If individuals with possible cash prescriptions (i.e., those described above) are assumed to have antipsychotic medication for all days from the last day covered to the end of the measurement period (i.e., 100% adherence), the PDC would be 71.6% (11,365/15,874).

SCENARIO 2. If individuals with possible cash prescriptions (i.e., those described above) are assumed to have antipsychotic medication for all days from the last day covered at the same proportion as the PDC calculated over the period from first to last claim in the measurement period (i.e., same adherence as the rest of the period), the PDC would be 71.5% (11,353/15,874).

The actual measure rate was 71.5% (11,348/15,874).

**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*)

**Missing Data**

Only 2 individuals (0.005%) in the overall measure denominator had one or more claims with missing days’ supply. This small number indicates that missing data do not pose a threat to the validity of the measure.

**Cash Prescriptions**

The actual measure rate was 71.5% (11,348/15,874). Therefore, the findings suggest that very little impact on measure rates would be expected from utilization of the cash discount program. In addition, since the most prevalent antipsychotic medications are not included in the commercial discount program due to their cost, it is unlikely that commercial discount programs will have an impact on measure rates in the near-term. Of note, this analysis is exploratory in nature and assumes that individuals were not switched to a drug on the commercial discount formulary, and if they were utilizing the discount program, they were obtaining all of their medications at a cash discount program. Additional limitations include prescriptions filled with other benefits (e.g., VA), and the extent to which this measure might underestimate antipsychotic use due to those factors is unknown.