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

**Measure Number** (*if previously endorsed*)**:** 0105

**Measure Title**: Antidepressant Medication Management

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

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | Efficiency |
| Structure |  |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

**2018 submission**

**2012 Submission**

This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via the Interactive Data Submission System (IDSS) portal.

The URL is: <http://www.ncqa.org/tabid/370/default.aspx>

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

**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: | other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

**2018 Submission**

Data for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 401 Medicare health plans, 226 Medicaid health plans, and 403 commercial health plans. The sample data included all Medicare, Medicaid and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Data for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 384 Medicare health plans, 184 Medicaid health plans, and 398 commercial health plans. The sample data included all Medicare, Medicaid and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

**2012 Submission**

The performance data for the past three years are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Data is summarized at the health plan level (i.e. the number of health plans). Data is stratified by year and product line (i.e. commercial, Medicare, Medicaid) The number of health plans submitting data for the Antidepressant Medication Management measure differs by product line; commercial – 385; Medicaid – 97; Medicare – 335.

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

**2018 Submission**

Patient population for measure score reliability testing: In 2016, HEDIS measures covered 114.2 million commercial health plan members, 47.0 million Medicaid members and 17.6 million Medicare beneficiaries. Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the population measured. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median number of eligible patients for this measure per plan |
| Commercial | 403 | 755 |
| Medicare | 401 | 322 |
| Medicaid | 226 | 1535 |

Patient population for Construct Validity Testing: In 2016, HEDIS measures covered 114.2 million commercial health plan members, 47.0 million Medicaid members and 17.6 million Medicare beneficiaries. Data is summarized at the health plan level. Data are stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the measured entities that include HEDIS data collection and the median eligible population for the measure across health plans.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of plans | Median number of eligible patients per plan |
| Commercial | 403 | 755 |
| Medicare | 401 | 322 |
| Medicaid | 226 | 1535 |

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

**2018 Submission**

N/A

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

**2018 Submission**

Measure performance was assessed by Medicaid, commercial and Medicare plan types.

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

**2a2. RELIABILITY TESTING**

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

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

**2018 Submission**

Reliability Testing of Performance Measure Score: same as below

**2012 submission**

NCQA estimates reliability with a beta-binomial model. The beta-binomial is a natural model for estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
**2018 Submission**

Beta-binomial statistic for each measure rate:

|  |  |  |  |
| --- | --- | --- | --- |
| Rate | Commercial | Medicare | Medicaid |
| Acute Phase | 0.97 | 0.97 | 0.99 |
| Continuation Phase | 0.97 | 0.97 | 0.99 |

**2012 submission**

Reliability for this measure as per the beta binomial model was calculated as 0.97 for the acute phase and 0.95 for the continuation phase.

**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?*)  
**2018 Submission**

Interpretation of measure score reliability testing for both measure rates: The testing suggests the measure has high reliability.

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

**2b1. VALIDITY TESTING**

**2b1.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*  
**2018 submission:**

We assessed both construct and face validity for this measure.

Method of testing construct validity:

We tested for construct validity by exploring whether Antidepressant Medication Management was correlated with Statin Therapy for Patients With Diabetes in Medicare, commercial, and Medicaid plans.

We hypothesized that organizations that perform well on the Antidepressant Medication Management measure should perform well on the Statin Therapy for Patients With Diabetes measure given that the measures are about health plans’ success improving adherence to medication treatment for chronic conditions.

To test these correlations, we used a Pearson correlation test. These tests estimate the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable. Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. Values between 0.3 and 0.7 indicate a moderate level. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the population is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

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

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

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

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

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQA’s Board of Directors will be included in the next HEDIS year and reported as first-year measures.

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

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

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

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

**2012 submission:**

Included below are the steps taken with all NCQA HEDIS measures, and more specific steps taken for the Antidepressant Medication Management measure.

NCQA uses a standardized process called the HEDIS measure life cycle to ensure the validity of measures.    
   
\*Step 1: Topic selection is the process of identifying measures that meet criteria consistent with the overall model for performance measurement. There is a huge universe of potential performance measures for future versions of HEDIS. The first step is identifying measures that meet formal criteria for further development.   
   
NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted.

\*Step 2: Development ensures that measures are fully defined and tested before the organization collects them. Field testing can involve parallel form testing using two different data sources (i.e. claims and paper records) or testing in several health plans. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase.

The Committee on Performance Measurement (CPM) uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.   
   
\*Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

\*Step 4: First-year data collection requires that organizations collect and report first-year measures and that those measures be available for audit. First-year measure results are not publicly reported and are not included in NCQA’s Quality Compass or in accreditation scoring.    
   
After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.   
   
\*Step 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be reported in Quality Compass and may be used for scoring in accreditation.    
   
Step 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reevaluated at least every three years.

---

AMM MEASURE DEVELOPMENT AND TESTING:

Step 1: NCQA developed the Antidepressant Medication Management measure to address the gap in care surrounding adherence to antidepressants for people diagnosed with major depression. NCQA’s Performance Measurement Department and the Behavioral Health MAP worked together to assess the most appropriate elements of this measure.  
  
Step 2: The measure was written, field-tested, and presented to the CPM and incorporated into HEDIS in 1998 for HEDIS 1999. After reviewing field test results, The CPM’s recommendation was to send the measure to public comment with a majority vote.  
  
Step 3: NCQA released the measure for Public Comment prior to publication in HEDIS. We received and responded to comments on this measure. Based on positive feedback, the CPM recommended moving this measure to first year data collection by a majority vote.   
  
Step 4: The Antidepressant Medication Management measure was introduced in HEDIS 1999. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure public reporting with a majority vote.

Step 5: The Antidepressant Medication Management measure was reevaluated in 2007 and 2012. The most recent field test data, from the re-evaluation in 2007 is presented below in section 2b2.3.

---

FIELD TESTING ANALYTIC METHOD:

For the field test, participating plans provided data beyond what would normally be necessary to compute this measure. They provided patient and pharmacy data from administrative data systems and medical records for the entire eligible population. Medical records accounted for 4.6 percent, 11.3 percent 50.3 percent of the total administrative claims and medical records submitted for the three plans. The reason for including certain information from both administrative sources and medical records, despite the measure being specified for administrative claims only, was to maximize the data found to help validate the measure. The 2007 field test was designed to answer several questions with respect to validity:

1. Is data available for identifying eligible patients?
2. Can the data identify negative-medication-history time periods with sufficient accuracy?
3. Does the length of the negative-medication history impact the denominator size?
4. Does the length of the continuous enrollment period impact the denominator size?
5. What percent of antidepressants prescribed are Tricyclic antidepressants?
6. What percent of diagnoses for major depression are accounted for by ICD-9 code: 311?

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

Statistical results of construct validity testing: The results in Table 1a showed that the Antidepressant Medication Management measure is significantly and positively correlated with the Statin Therapy for Patients With Diabetes measure and the correlation was moderate (the correlation coefficients are higher than 0.3).

**Table 1a. Correlations between Antidepressant Medication Management Other Quality Measures in Medicaid Plans – HEDIS 2017**

|  |  |
| --- | --- |
| **Pearson Correlation Coefficients** | Statin Therapy for Patients With Diabetes (Statin Adherence Indicator: Members who remained on a statin medication of any intensity for at least  80% of the treatment period) |
| Antidepressant Medication Management – Acute Phase | 0.50 |
| Antidepressant Medication Management – Continuation Phase | 0.49 |

Note: p<0.0001

The results in Table 1b and 1c indicate that there is a strong positive relationship between the Antidepressant Medication Management measure and the Statin Therapy for Patients With Diabetes (Statin coverage rate) measure in commercial and Medicare plans. This relationship is statistically significant (p<0.0001).

**Table 1b. Correlations between the Antidepressant Medication Management and Statin Therapy for Patients With Diabetes measures in Commercial Plans – HEDIS 2017**

|  |  |
| --- | --- |
| **Pearson Correlation Coefficients** | Statin Therapy for Patients With Diabetes (Statin Adherence Indicator: Members who remained on a statin medication of any intensity for at least  80% of the treatment period) |
| Antidepressant Medication Management – Acute Phase | 0.69 |
| Antidepressant Medication Management – Continuation Phase | 0.69 |

Note: p<0.0001

**Table 1c. Correlations between the Antidepressant Medication Management and Statin Therapy for Patients With Diabetes measures in Medicare Plans – HEDIS 2017**

|  |  |
| --- | --- |
| **Pearson Correlation Coefficients** | Statin Therapy for Patients With Diabetes (Statin Adherence Indicator: Members who remained on a statin medication of any intensity for at least  80% of the treatment period) |
| Antidepressant Medication Management – Acute Phase | 0.56 |
| Antidepressant Medication Management – Continuation Phase | 0.60 |

Note: p<0.0001

Results of face validity assessment:

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

**2012 submission:**

NCQA field tested the measure in 1998 and again in 2007.

NCQA developed the measure through the Robert Wood Johnson Chronic Disease Grant. The field testing in 1996 included two health plans. The field test design had two goals:

1. Find out if the measure should focus on appropriate dosing of antidepressants or adherence.
2. See a relationship between adherence and depression relapse.

The pilot testing results demonstrated that continuation of therapy was a more feasible approach to measure appropriate pharmacotherapy for people with major depression. Ninety percent of patients receiving continuation therapy were receiving effective therapeutic doses, which left little room for performance improvement. Patients who remained on an effective therapeutic dose of a recommended antidepressant were significantly more likely to experience symptom resolution than patients who discontinue their medication prematurely. Therefore, NCQA’s expert panel recommended and the CPM voted to include the measure in HEDIS 1999.

The results from the most recent field test demonstrated high levels of concordance between the performance rates and denominator percentages of the field test and our HEDIS data. The field test data demonstrates that the specifications are highly reliable and accurate in identifying patients with major depression and those who were prescribed an antidepressant. Plans were able to calculate the negative medication histories and correctly follow the continuous enrollment criteria. The current measure’s intent is to focus on new treatment episodes of depression; therefore, the current measure does not include the negative diagnosis history. The testing results summarized below exclude the negative diagnosis history results for that reason.

Question 1.

* For the three plans, the average percent of the eligible population with a major depression diagnosis was 9.5 percent.

Question 2:

* Through both administrative claims and medical record data, plans can find the length of time prior to the index prescription date that a member was prescribed an antidepressant.

Question 3:

* Health plans were concerned that 90 days is not sufficient to identify people currently on an antidepressant. Those concerns contend that extending the period would more accurately exclude people being treated with antidepressants prior to the index prescription date. The current negative medication history is 90 days, which aligns with the continuous enrollment period of 90 days. If the negative medication history was increased, to address this concern, an additional 4 percent of the eligible population would be excluded. NCQA’s experts felt that it was unnecessary to exclude more patients, because most prescriptions for antidepressants are for 90 days or less. Therefore, the measure accurately excludes people that are not newly treated with antidepressants.

Question 4:

* If the continuous enrollment period was extended to align with any extension in the negative medication history, a higher percent of the eligible population would be excluded. If it was increased to 120 days, an additional 11.5 percent of the eligible population would be excluded.

Question 5:

* Health plans were concerned that including Tricyclic antidepressants (TCAs) in the measure will produce inaccuracies, because often times TCAs are not a first line pharmacy option for major depression. The field test data shows that TCAs only account for on average 2.25 percent of the antidepressants prescribed. Therefore, our expert panels advised NCQA to keep the TCAs in the measure as a treatment option.

Question 6:

* Health plans were concerned that ICD-9 code 311 is a “catch-all” for major depression, and is inappropriately used by health plans. The field test data shows that code 311 accounts for between 31 percent and 41 percent of the diagnosis codes used to identify Major Depression. Because of its common use, and because the measure also includes a prescription for an antidepressant, which helps confirm the major depression diagnosis, NCQA’s expert panels advised NCQA to keep the code in the measure.

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)  
**2018 Submission**

Interpretation of construct validity testing: The Antidepressant Medication Management measure was positively correlated with Statin Therapy for Patients With Diabetes (0.49-0.69), suggesting they represent the same underlying quality construct of quality of care. Therefore, health plans that performed well on antidepressant medication management should also provide good statin therapy for patients with diabetes, which indicates the measure has strong construct validity.

These results suggest that the Antidepressant Medication Management measure is a valid measure of a plan’s quality of adhering to medications for chronic diseases.

Interpretation of systematic assessment of face validity: These results indicate the technical expert panel showed good agreement that the measures as specified will accurately differentiate quality across providers. Our interpretation of these results is that this measure has sufficient face validity.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**2b2.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)  
 Testing was not performed for exclusions.

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)  
Testing was not performed for exclusions.

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)  
Testing was not performed for exclusions.

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

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**   
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b4***](#section2b5)***.***

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

**2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

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

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**2012 submission**

The inter-quartile range was calculated to determine the variability of performance on the measure. The inter-quartile range provides a measure of the dispersion of performance.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
**2018 Submission**

HEDIS 2017 Variation in Performance across Health Plans for Acute Phase

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Commercial | 1,958 | 67.5 | 6.5 | 58.6 | 64.0 | 67.5 | 71.8 | 75.7 | 7.8 | <0.001 |
| Medicare | 1,010 | 70.2 | 8.8 | 59.4 | 64.8 | 70.7 | 75.9 | 80.3 | 11.1 | <0.001 |
| Medicaid | 2,301 | 53.2 | 8.8 | 44.5 | 48.2 | 51.9 | 57.5 | 64.2 | 9.3 | <0.001 |

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

IQR: Interquartile range

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile. P-values are less than 0.05.

HEDIS 2017 Variation in Performance across Health Plans for Continuation Phase

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Commercial | 1,958 | 51.8 | 6.8 | 43.4 | 47.6 | 51.5 | 56.0 | 60.4 | 8.4 | <0.001 |
| Medicare | 1,010 | 55.5 | 10.3 | 42.1 | 48.9 | 55.6 | 61.2 | 67.5 | 12.3 | <0.001 |
| Medicaid | 2,301 | 38.0 | 9.4 | 29.1 | 32.6 | 36.3 | 41.6 | 50.4 | 9.0 | <0.001 |

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

IQR: Interquartile range

p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile. P-values are less than 0.05.

**2012 submission**

There has been slow and steady improvement in performance in commercial, Medicare and Medicaid product lines over the last six years. Rates have gradually increased across means and percentiles at about the same rate. In general, rates are higher for the acute phase than the continuation phase, and higher in Medicare. Over the last three years, the number of plans reporting in the Medicare and Medicaid product lines has increased (close to 100 plans for Medicare), and dropped slightly in commercial. The data illustrates continued gaps in performance.

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)  
**2018 Submission**

The difference between the 25th and 75th percentile is statistically significant for both rates in all product lines.

In commercial plans, there is a 7.8 percentage point gap between 25th and 75th percentile plans for the acute phase rate. This gap represents an average 153 more patients who have remained on an antidepressant medication for at least 84 days (12 weeks) compared to low performing plans (estimated from average health plan eligible population). For the continuation phase rate, there is a 8.4 percentage point gap between 25th and 75th percentile plans. This gap represents an average 164 more patients who have remained on an antidepressant medication for at least 180 days (6 months) compared to low performing plans (estimated from average health plan eligible population).

In Medicare plans, there is a 11.1 percentage point gap between 25th and 75th percentile plans for the acute phase rate. This gap represents an average 112 more patients who have remained on an antidepressant medication for at least 84 days (12 weeks) compared to low performing plans (estimated from average health plan eligible population). For the continuation phase rate, there is a 12.3 percentage point gap between 25th and 75th percentile plans. This gap represents an average 124 more patients who have remained on an antidepressant medication for at least 180 days (6 months) compared to low performing plans (estimated from average health plan eligible population).

In Medicaid plans, there is a 9.3 percentage point gap between 25th and 75th percentile plans for the acute phase rate. This gap represents an average 214 more patients that have who remained on an antidepressant medication for at least 84 days (12 weeks) compared to low performing plans (estimated from average health plan eligible population). For the continuation phase rate, there is a 9.0 percentage point gap between 25th and 75th percentile plans. This gap represents an average 207 more patients that have who remained on an antidepressant medication for at least 180 days (6 months) compared to low performing plans (estimated from average health plan eligible population).

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

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

**2018 Submission**

This measure is collected with a complete sample.

**2012 submission**

This measure is precisely specified using the administrative data collection method. This measure has detailed, precise specifications that clearly define the numerator, denominator, data sources, allowable values, methods of measurement and reporting.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)  
**2018 Submission**

This measure is collected with a complete sample.

**2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

**2018 Submission**

This measure is collected with a complete sample.