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

**Measure Number** (*if previously endorsed*)**:** 0108

**Measure Title**: Follow-Up Care for Children Prescribed ADHD Medications

**Date of Submission**: 1/6/2020

**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*).

**2020 Submission**

N/A

**2016 Submission**

**1.3. What are the dates of the data used in testing**? Click here to enter date range

**2020 Submission**

Testing of measure score reliability and validity was performed using data from calendar years 2016, 2017, and 2018.

**2016 Submission**

2014-2016

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.20*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

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

**2020 Submission**

This measure assesses whether children (age 6-12) newly prescribed attention-deficit/hyperactivity disorder (ADHD) medication receive at least three follow-up visits within a 10-month period, one of which occurring within 30 days of the first ADHD medication dispensed. This measure includes patients enrolled in commercial and Medicaid health plans. The first rate is for the Initiation Phase, assessing the percentage of members who had one follow-up visit with a prescribing practitioner within the 30 days of the first ADHD medication dispensed. The second rate is for the Continuation and Maintenance (C&M) Phase, assessing the percentage of members who remained on the medication for at least 210 days and had at least two follow-up visits with a practitioner within 9 months after the Initiation phase. The intended use of the measure is to assess the quality of care in health plans across the general child population. As required by the specified level of accountability, we conducted a field test with health plans to assess scientific acceptability, usability and feasibility and have subsequently gathered audited data from a large number of health plans.

Sample for measure score reliability testing and construct validity testing: The measure score reliability was calculated from HEDIS data that included 319 commercial health plans and 183 Medicaid health plans for the Initiation Phase and 205 commercial health plans and 169 Medicaid health plans for the Continuation and Maintenance Phase. The sample included all commercial and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

**2016 Submission**

**Update: MEASURE SCORE RELIABILITY TESTING**

The measure score reliability was calculated from HEDIS data that included 344 Commercial health plans and 180 Medicaid health plans for the Initiation Phase and 210 Commercial health plans and 151 Medicaid health plans for the Continuation and Maintenance Phase. The sample included all Commercial and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

**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*)   
**2020 Submission**

For HEDIS 2019 (calendar year 2018), HEDIS measures covered 116 million commercial health plan members and 54 million Medicaid enrollees. Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included in HEDIS data collection and the average eligible population for the measure across health plans.

Table 1. Mean eligible population for the *Follow-Up Care for Children Prescribed ADHD Medication* measure by plan type, Initiation Phase, calendar year 2018 data

|  |  |  |
| --- | --- | --- |
| **Product Type** | **Number of Plans** | **Mean number of eligible members per plan** |
| Commercial | 319 | 417 |
| Medicaid | 183 | 1,466 |

Table 2. Mean eligible population for the *Follow-Up Care for Children Prescribed ADHD Medication* measure by plan type, Continuation and Maintenance Phase, calendar year 2018 data

|  |  |  |
| --- | --- | --- |
| **Product Type** | **Number of Plans** | **Mean number of eligible members per plan** |
| Commercial | 319 | 162 |
| Medicaid | 183 | 340 |

**2016 Submission**

**2016 Update: MEASURE SCORE RELIABILITY TESTING**

Patient data set for measure score reliability testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare beneficiaries. This measure applies to commercial and Medicaid plans. Data are summarized at the health plan level and stratified by product line. Below is a description of the sample, including number of health plans included and the median eligible population for the measure across health plans.

INITIATION PHASE

|  |  |  |
| --- | --- | --- |
| **Product Line** | **Number of Plans** | **Mean number of eligible patients per plan** |
| Commercial | 344 | 397 |
| Medicaid | 180 | 1,160 |

*CONTINUATION AND MAINTENANCE PHASE*

|  |  |  |
| --- | --- | --- |
| ***Product Line*** | **Number of Plans** | **Mean number of eligible patients per plan** |
| Commercial | 210 | 163 |
| Medicaid | 151 | 320 |

**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**.

**2020 Submission**

No differences in the data used for reliability and construct validity testing.

**2016 Submission**

**2016 Update: MEASURE SCORE RELIABILITY TESTING**

Reliability of the measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

**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.

**2020 Submission**

Social risk factor data were not available in reported results. This measure is specified to be reported separately by Medicaid and commercial plan types, which serves as a proxy for income and other socioeconomic factors.

**2016 Submission**

2016 Update: Measure performance was assessed by commercial, Medicaid 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*)

**2020 Submission**

Reliability testing of performance measure score

We utilized the Beta-binomial model (Adams 2009) to assess how well one can confidently distinguish the performance of one accountable entity from another. Conceptually, the Beta-binomial model is the ratio of signal to noise. The signal is the proportion of the variability in measured performance that can be explained by real differences in performance. The Beta-binomial model is an appropriate model when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS measures. Reliability scores range from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (i.e., noise), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

In addition to the point estimate of signal-to-noise reliability, NCQA will also provide the standard error and 95% confidence interval (95% CI) by June 2, 2020. NCQA will include a summary of the methodology that was used to estimate the standard error and 95% CI.

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

**2016 Submission**

**2016 Update: Method for measure score reliability testing**

We used the beta binomial method as described below in our 2013 submission.

**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*)

**2020 Submission**

Table 3 shows the estimated signal-to-noise reliability for each indicator.

Table 3. Estimated Signal-to-Noise Reliability for *Follow-Up Care for Children Prescribed ADHD Medication,* for Commercial and Medicaid Plans, calendar year 2018 data

|  |  |  |
| --- | --- | --- |
| **Measure Rate** | **Signal-to-Noise Reliability** | |
| **Commercial** | **Medicaid** |
| Initiation Phase | 0.88 | 0.98 |
| Continuation and Maintenance Phase | 0.74 | 0.94 |

\* NCQA will provide the standard error and 95% CI for signal-to-noise reliability by June 2, 2020.

**2016 Submission**

**2016 Update: MEASURE SCORE RELIABILITY**

Beta-Binomial Statistic For Each Measure Rate: Mean Reliability

|  |  |  |
| --- | --- | --- |
| *Rate* | *Commercial* | *Medicaid* |
| Initiation Phase | 0.90 | 0.98 | |
| Continuation and Maintenance Phase | 0.75 | 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?*)

**2020 Submission**

The signal-to-noise reliability estimates are greater than 0.7, indicating the measure has very good reliability and provides confidence that one can distinguish the performance of one plan from another.

**2016 Submission**

**2016 Update: interpretation of results for measure score reliability testing**

Beta binomial testing for this measure suggests the two rates (Initiation and Continuation and Maintenance) within this measure have strong reliability for commercial (0.90, 0.75) and Medicaid (0.98, 0.95) health plans.

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

**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)*  
**2020 Submission**

We assessed construct validity and face validity for this measure.

Method of testing construct validity

We tested for construct validity by exploring the following:

* Are the individual rates within the *Follow-Up Care for Children Prescribed ADHD Medication* measure correlated with one another.
* Is *Follow-Up Care for Children Prescribed ADHD Medication* correlated with the following HEDIS measure, *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics*, which assesses the proportion of children and adolescents without a primary indication who had a new prescription for an antipsychotic medication and had documentation of psychosocial care as first-line treatment.

We hypothesized that rates within the *Follow-Up Care for Children Prescribed ADHD Medication* measure would be highly correlated, and that organizations that perform well on *Follow-Up Care for Children Prescribed ADHD Medication* should perform well on the other measure, *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics*, given that they address the same or similar child populations and similar behavioral health conditions. To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two variables. The magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

Method of assessing face validity

NCQA develops measures using a standardized process. For new measures, face validity is assessed at various steps as described below.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical measurement advisory 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, which is vetted by the 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. At this step, face validity is systematically determined by the CPM, which 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 about proposed new measures. Public comment offers an opportunity to assess the validity, feasibility, importance and other attributes of a measure from a wider audience. For this measure, a majority of public comment respondents supported the measure. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM.  Face validity is then again systematically assessed by the CPM. The CPM reviews all comments before making a final decision and votes to recommend approval of new measures for HEDIS. NCQA’s Board of Directors then approves new measures.

**2016 Submission**

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
**2020 Submission**

Statistical results of construct validity testing

Table 4a. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up Care for Children Prescribed ADHD Medication* Performance Scores Within Measure – **Commercial** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | Initiation |
| Continuation and Maintenance | 0.78\* |

\*Significant at p<0.05

Table 4b. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up Care for Children Prescribed ADHD Medication* Performance Scores Within Measure – **Medicaid** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | Initiation |
| Continuation and Maintenance | 0.89\* |

\*Significant at p<0.05

Table 5a. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up Care for Children Prescribed ADHD Medication* and *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* Measure Performance Scores – **Commercial** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* |
| *Follow-Up Care for Children Prescribed ADHD Medication* |  |
| Initiation | 0.26\* |
| Continuation and Maintenance | 0.14 |

\*Significant at p<0.05  

Table 5b. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up Care for Children Prescribed ADHD Medication* and *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* Measure Performance Scores – **Medicaid** Plans, calendar year 2018 data

|  |  |
| --- | --- |
| **Rate** | **Correlation Coefficient** |
|  | *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* |
| *Follow-Up Care for Children Prescribed ADHD Medication* |  |
| Initiation | 0.31\* |
| Continuation and Maintenance | 0.30\* |

\*Significant at p<0.05

Results of face validity assessment

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

**2016 Submission**

**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?*)  
**2020 Submission**

Interpretation of construct validity testing    
Correlations between individual rates within the *Follow-Up Care for Children Prescribed ADHD Medication* measure were strong (Tables 4a, 4b) across product lines. In the commercial product line, correlations between the *Follow-Up Care for Children Prescribed ADHD Medication* and the *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* measure rates (Table 5a) were weak to moderate. In the Medicaid product line, correlations between the *Follow-Up Care for Children Prescribed ADHD Medication* and the *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* measure rates (Table 5b) were strong. Plans with higher rates on *Follow-Up Care for Children Prescribed ADHD Medication* tend to also have higher rates on the *Use of First-Line Psychosocial Care for Children and Adolescents on Antipsychotics* measure. The results indicate that the *Follow-Up Care for Children Prescribed ADHD Medication* measure has good validity.

Interpretation of systematic assessment of face validity

 The multi-stakeholder advisory panels concluded the measures had good face validity.

**2016 Submission**

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

**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*)

**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*)

**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*)

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

**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)***.***

**2020 Submission**

N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure.

**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)*   
**2020 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.

**2016 Submission**

**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*)  
**2020 Submission**

Table 7. Calendar year 2018 Variation in Performance Across Health Plans

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Product Line** | **Rate** | **Avg. EP** | **Avg.**  **(%)** | **SD**  **(%)** | **10th**  **(%)** | **25th**  **(%)** | **50th**  **(%)** | **75th**  **(%)** | **90th**  **(%)** | **IQR**  **(%)** | **p-value** |
| Commercial | Initiation | 417 | 40.0 | 8.3 | 29.8 | 35.0 | 40.0 | 44.3 | 50.4 | 15.4 | 0.0000 |
| Continuation & Maintenance | 162 | 48.0 | 9.3 | 36.2 | 43.2 | 47.9 | 53.0 | 60.0 | 9.8 | 0.0017 |
| Medicaid | Initiation | 1,466 | 44.2 | 9.7 | 33.9 | 37.9 | 43.4 | 49.9 | 56.6 | 12.0 | 0.0000 |
| Continuation & Maintenance | 340 | 54.6 | 12.0 | 39.0 | 46.4 | 55.5 | 62.7 | 71.2 | 16.3 | 0.0000 |

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.

**2016 Submission**

**2016 Update: Ability to identify statistically significant/meaningful differences**

HEDIS 2016 Variation in Performance across Health Plans- Commercial

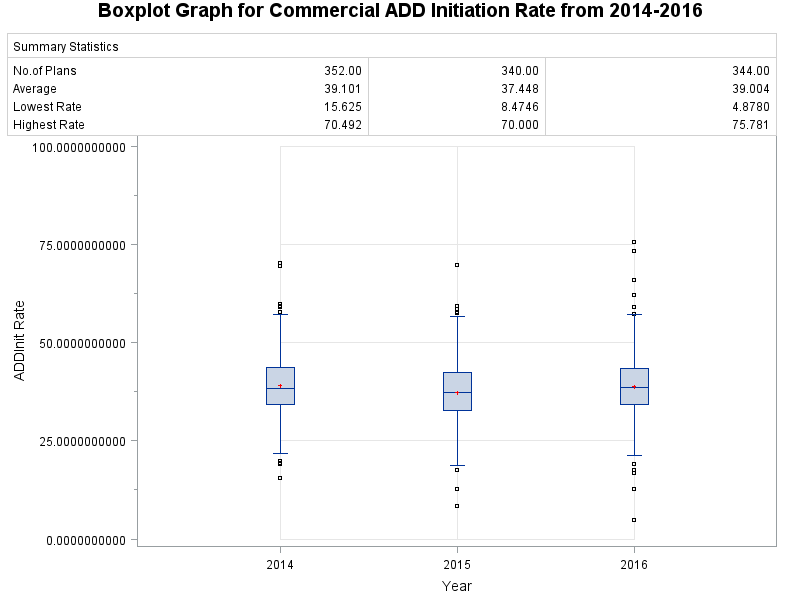
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Product Line | Rate | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Commercial | Initiation | 397 | 39.0% | 8.6% | 29.1% | 34.3% | 38.6% | 43.5% | 50.2% | 9.2% | <0.001 |
| C&M | 163 | 46.8% | 9.3% | 35.6% | 40.7% | 46.4% | 52.3% | 57.3% | 11.6% | 0.002 |
| Medicaid | Initiation | 1,160 | 42.2% | 11.0% | 28.8% | 34.2% | 42.2% | 49.6% | 55.5% | 15.4% | <0.001 |
| C&M | 320 | 50.9% | 13.3% | 34.0% | 40.9% | 52.5% | 62.5% | 67.2% | 21.6% | <0.001 |

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

IQR: Interquartile range

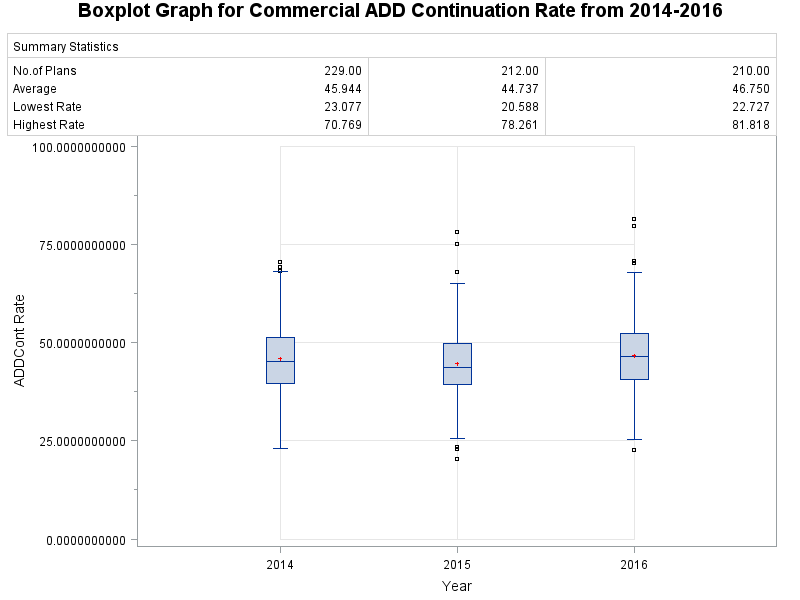
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

**Figure 1a. *Follow-up Care for Children Prescribed ADHD Medication -* *Initiation Phase*: Commercial Plans 2014-2016**

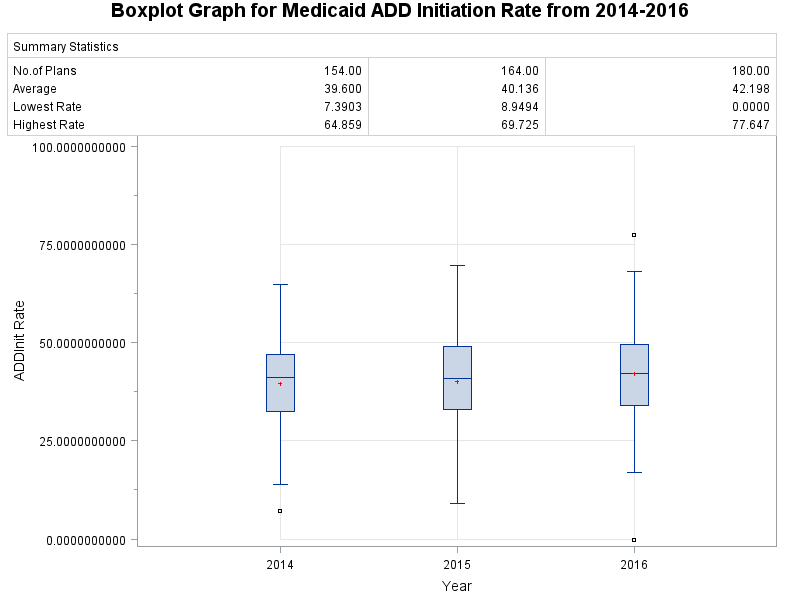


Page Break

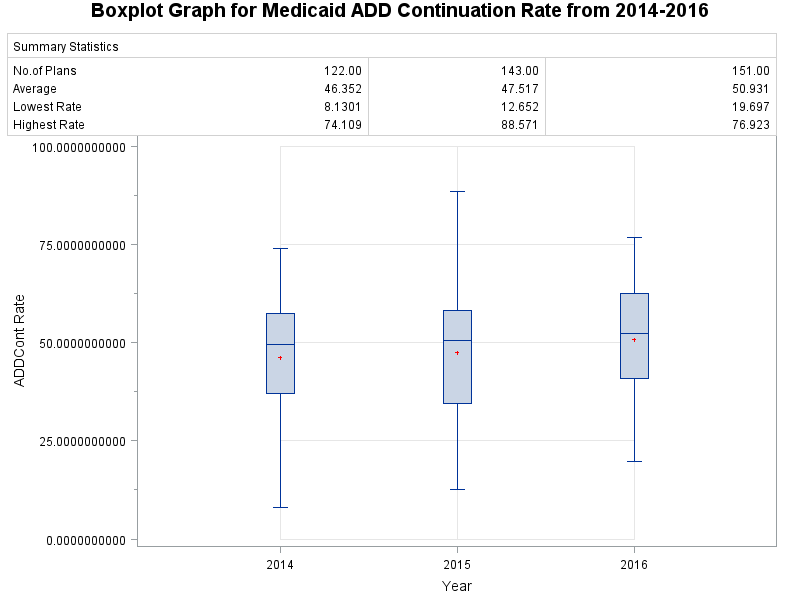
**Figure 1b. *Follow-up Care for Children Prescribed ADHD Medication –* *C&M Phase*: Commercial Plans 2014-2016**



**Figure 2a. *Follow-up Care for Children Prescribed ADHD Medication -* *Initiation Phase*: Medicaid Plans 2014-2016**



**Figure 2b. *Follow-up Care for Children Prescribed ADHD Medication –* *C&M Phase*: Medicaid Plans 2014-2016**



**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?*)  
**2020 Submission**

The results above indicate there is a 9.8-16.3% gap in performance between the 25th and 75th performing plans. For all product lines and rates, the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for Medicaid plans rates, which show a 12-16.3 percentage point gap between 25th and 75th percentile plans. This gap represents on average 176 and 54 more children receiving Initiation Phase and Continuation and Maintenance Phase care respectively in high performing Medicaid plans compared to low performing plans

**2016 Submission**

**2016 Update: Interpretation of Ability to identify statistically significant/meaningful differences**

The results above indicate there is a 9-22% gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the Medicaid health plans which show a 15.4-21.6 percentage point gap between 25th and 75th percentile plans. This gap represents on average 179  children in the Initiation Phase and 69 children in the Continuation and Maintenance Phase in high performing Medicaid plans compared to low performing plans (estimated from average health plan eligible population). Additionally, on average, plans in the 90th percentile performed approximately 21 percentage points better than plans in the 10th percentile in the commercial product line. In the Medicaid product line, on average, plans in the 90th percentile performed approximately 30 percentage points better than plans in the 10th percentile. Overall, these results suggest there are meaningful differences in performance and there is an opportunity for improvement.

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

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

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

**2020 Submission**

This measure has only one set of specifications.

**2016 Submission**

**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*)  
**2020 Submission**

HEDIS measures apply to enrolled members in a health plan, and NCQA has a rigorous audit process to ensure the eligible population and numerator events for each measure are correctly identified and reported. The audit process is designed to verify primary data sources used to populate measures and ensure specifications are correctly implemented.

The HEDIS Compliance Audit addresses the following functions:

- Information practices and control procedures

- Sampling methods and procedures

- Data integrity

- Compliance with HEDIS specifications

- Analytic file production

- Reporting and documentation

**2016 Submission**

**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*)  
**2020 Submission**

HEDIS addresses missing data in a structured way through its audit process. HEDIS measures apply to enrolled members in a health plan, and NCQA-certified auditors use standard audit methodologies to assess whether data sources are missing data. If a data source is found to be missing data, and the issues cannot be rectified, the auditor will assign a “materially biased” designation to the measure for that reporting plan, and the rate will not be used. Once measures are added to HEDIS, NCQA conducts a first-year analysis to assess the measure’s feasibility once widely implemented in the field. This analysis includes an assessment of how many plans report valid rates vs. rates that are materially biased. These considerations are weighed in the deliberation process before measures are approved for public reporting.

**2016 Submission**

**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*)

**2020 Submission**

This measure goes through the NCQA audit process each year to identify potential errors or bias in results. Only performances rates that have been reviewed and determined not to be “materially biased” are reported and used.

**2016 Submission**