NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Antipsychotic Use in Children Under 5 Years Old
Date of Submission: 1/16/2014
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	X Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decision-making) 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]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14],[15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

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

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
X administrative claims	X administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
other:	other:

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

The dataset used for testing of this measure consisted of prescription drug claims data from the 2007 Medicaid MAX File for Medicaid patients in 44 states and the District of Columbia. These data were obtained from the Centers for Medicare and Medicaid Services (CMS) for the following states: Alabama, Arkansas, Arizona, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Iowa, Idaho, Illinois, Indiana, Kansas, Kentucky, Louisiana, Massachusetts, Maryland, Michigan, Minnesota, Missouri, Mississippi, North Carolina, Nebraska, New Hampshire, New Jersey, New Mexico, Nevada, New York, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Utah, Virginia, Vermont, Washington, Wisconsin, and West Virginia.

The multi-state prescription drug data included claims from January 1, 2007 to December 31, 2007. The data from this time period were the most complete recent data available to PQA at the time of testing.

The measure was further tested using more recent prescription claims data from Mississippi's Medicaid program. That time period was January 1, 2012 to December 31, 2012.

1.3. What are the dates of the data used in testing?

The majority of testing used multi-state Medicaid prescription claims data from January 1, 2007 to December 31, 2007. The data from this time period were the most complete recent data available at the time of testing. Testing also included prescription claims data from one state's Medicaid data from January 1, 2012 to December 31, 2012.

It is important to note that the use of antipsychotics in pediatric populations is increasing. A recent publication examined the prevalence of antipsychotic medications in pediatric Medicaid-eligible patients over a 10-year period and found an increase in usage from 1.2% to 3.2% over that time period. The Wall Street Journal (WSJ) cited a Mathematica analysis that found a three-fold increase in Medicaid prescriptions from 1999 to 2008 for antipsychotic medications in patients under 20 years of age. The WSJ also cited Medicaid representatives who reported that in 2008, 19,045 children under the age of 5 were prescribed antipsychotics through Medicaid, up from 7,759 in 1999.

In recent years, while some Medicaid programs have focused on this concern and because of their scrutiny, some have seen a plateau and even a slight decrease in rates of antipsychotic use in children, the problem overall is still of much concern and in need of focus.

- Lagnado L. U.S. Probes Use of Antipsychotic Drugs on Children: Federal health officials are reviewing antipsychotic drug use on children in the Medicaid system. Wall Street Journal. 2013 Aug 11 [accessed 2013 Aug 11]. Available from: http://online.wsj.com/news/articles/SB10001424127887323477604578654130865747470
- Pringsheim T, Lam D, Patten SB. The pharmacoepidemiology of antipsychotic medications for Canadian children and adolescents: 2005-2009. J Child Adolesc Psychopharmacol. 2011 Dec;21(6):537-43. doi: 10.1089/cap.2010.0145. Epub 2011 Dec 2. Available from: http://online.liebertpub.com/doi/abs/10.1089/cap.2010.0145
- Zito JM, Burcu M, Ibe A, Safer DJ, Magder LS. Antipsychotic use by medicaid-insured youths: impact of eligibility and psychiatric diagnosis across a decade. Psychiatr Serv. 2013 Mar 1;64(3):223-9. doi: 10.1176/appi.ps.201200081. Available from: http://ps.psychiatryonline.org/article.aspx?articleid=1486122

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
X health plan	X health plan
X other: State-level Medicaid	X other: State-level Medicaid

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*)

The testing and analysis included forty-five (45) measured entities. The measured entities were state-level Medicaid prescription drug programs. The included states and sizes of the population are as follows in 1.6. (Note: N = the population under 5 years old, i.e. the measure denominator.)

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

The included states and sizes of the population are as follows: (Note: N = the population under 5 years old, i.e. the measure denominator.)

2007 Medicaid Claims Data

	N	% of All Benes		N	% of All Benes
State	**(< 5 years)**	**Using AP Med**	**State**	**(< 5 years)**	**Using AP Med**
AL	184,094	0.21%	**MS**	166,277	0.13%
AR	151,956	0.35%	**NC**	405,779	0.13%
AZ	298,135	0.09%	**NE**	69,097	0.32%
CA	1,340,065	0.03%	**NH**	29,286	0.12%
CO	151,963	0.09%	**NJ**	198,878	0.08%
CT	93,099	0.00%	**NM**	103,719	0.08%
DC	28,985	0.01%	**NV**	51,112	0.14%
DE	34,142	0.10%	**NY**	719,505	0.08%
FL	579,748	0.09%	**OH**	422,266	0.00%
GA	444,040	0.11%	**OK**	194,234	0.15%
HI	43,124	0.01%	**OR**	114,176	0.08%
IA	95,421	0.20%	**PA**	364,165	0.08%
ID	62,603	0.12%	**RI**	35,981	0.12%
IL	529,668	0.08%	**SC**	193,881	0.08%
IN	249,093	0.22%	**TN**	271,788	0.09%
KS	94,440	0.24%	**TX**	1,270,853	0.19%
KY	169,182	0.26%	**UT**	80,888	0.08%
LA	229,715	0.16%	**VA**	209,409	0.11%
MA	180,272	0.07%	**VT**	21,474	0.13%
MD	188,369	0.11%	**WA**	241,056	0.05%
MI	369,539	0.14%	**WI**	181,700	0.14%
MN	140,723	0.10%	**WV**	71,434	0.21%
MO	226,899	0.21%			

2012 Mississippi Medicaid Claims Data

	N	% of All Benes
State	**(< 5 years)**	**Using AP Med**
Mississippi	167,482	0.11%

A total of 11,302,233 Medicaid patients under the age of 5 were included in the testing and analysis of Medicaid data from the calendar year 2007. The data obtained in the study can be broken down by state in terms of patient demographics, age, managed care status, race, and percentage of foster care.

Characteristics of the total population include:

- Gender: 48.8% of the patients are females with 51.2% being males;
- Age: 56.2% of the population were 2 years of age or younger, 15.9% were 3 years of age, 15.2% were 4 years of age, and 12.7% were 5 years of age at the end of the observation period;
- Race: 36.5% were Caucasian, 22.4% were African American, 29.5% were Hispanic, and 11.5% were classified as other;
- Managed Care Status: 8.9% of the patients were enrolled in a fee-for-service plan only (FFS), 64.6% were enrolled in a managed care plan only (MAN), and 26.5% were enrolled in mixed plans; *and*
- Foster Care: 2% of the patients were in the foster care system.

Total Population	Gender		Age at End of Observation Period			
	Female	Male	<=2 yrs	3 yrs	4 yrs	5 yrs
11,302,233	48.8%	51.2%	56.2%	15.9%	15.2%	12.7%

Race				Managed Care Status			Foster Child	
White	Black	Hispanic	Other	FFS only	MAN only	Mixed	Number	%
6.5%	22.4%	29.5%	11.5%	8.9%	64.6%	26.5%	275,451	2.0%

A total of 167,482 Medicaid patients under the age of 5 were included in the Mississippi 2012 testing and analysis.

Characteristics of the population include:

- Gender: 49.3% of the patients are female with 50.7% being male;
- Age: 51.6% of the population were 2 years of age or younger, 16.6% were 3 years of age, 16.8% were 4 years of age, and 15.1% were 5 years of age at the end of the observation period;
- Race: 35.2% were Caucasian, 56.3% were African American, 4.7% were Hispanic, and 3.7% were classified as other;
- Managed Care Status: 86.1% of the patients were enrolled in a fee-for-service plan only, 1.3% were enrolled in a managed care plan only, and 12.6% were enrolled in mixed plans; *and*
- Foster Care: 0.8% of the patients were in the foster care system.

Mississippi Eligible Population	Gender		Age at End of Observation Period			
	Female	Male	<=2 yrs	3 yrs	4 yrs	5 yrs
167,482	49.3%	50.7%	51.6%	16.6%	16.8%	15.1%

Race				Managed Care Status			Foster Child	
White	Black	Hispanic	Other	FFS only	MAN only	Mixed	Number	%
35.2%	56.3%	4.7%	3.7%	86.1%	1.3%	12.6%	1,326	0.8%

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.

N/A

2a2. RELIABILITY TESTING
<u>Note</u>*: 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*)
X **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
X **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*)

Critical data elements:

The critical data elements used in this measure include: age, prescription medications filled, and plan enrollment. This measure uses pharmacy claims data and beneficiary enrollment data, which include these data elements. Enrollment was determined from beneficiary summary files. Monthly enrollment data were available for all beneficiaries.

Birth date was available in the beneficiary summary files and was included in each prescription record in the MAX-RX files. Reliability of age information was assessed by examining the percentage of beneficiaries for which data were not available.

The measure only requires one month of enrollment for a beneficiary to be eligible for inclusion in the denominator. Monthly enrollment data were available for all beneficiaries and were considered to be reliable since this is the data used by the state Medicaid agencies to determine coverage and without coverage, a beneficiary could not have a prescription claim paid.

Reliability of pharmacy claims data were not directly assessed for this measure, but literature has shown that prescription claims data are reliable and valid measures of medication use.

- Kirking DM, Ammann MA, Harrington CA. Comparison of Medical Records and Prescription Claims Files in Documenting Prescription Medication Therapy," J Pharmacoepi. 1996, 5(1):3-15.
- Choo PW, Rand CS, Inue TS, et al. Validation of patient reports, automated pharmacy records, and pill counts with electronic monitoring of adherence to antihypertensive therapy. Med Care 1999;37:846-57.
- Kwon A, Bungay KM, Pei Y, et al. Antidepressant use: concordance between self-report and claims records. Med Care 2003;41:368-74.
- Saunders K, Simon G, Bush T, Grothaus L. Validation of pharmacy records in drug exposure assessment. J Clin Epidemiol 1997;50:619-25.

Performance measure score:

An hierarchical multivariable logistic regression analysis was performed to control the effects of potential confounding variables while measuring the effect associated with state. Multilevel models are appropriate for this data, considering the hierarchical structuring of the data, where patients are clustered within states. Performance on the measure was adjusted for beneficiary age and race in the model. The hierarchical approach treats states as random effects and allows adjustment for within-state correlation in the process measure.

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*)

Age data were not available for only 1.8% of beneficiaries in the 45 states for the national analysis. 25 states had less than 1.0% missing, 9 states had 1.0-1.9% missing, and 11 states had 2.0% or more missing.

In the hierarchical logistic regression model with random-intercept, the state-level variance component was estimated to be 1.0436 (SE: 0.2563). Testing the null hypothesis of no random effects using a likelihood ratio test based on residual pseudo-likelihood yielded a chi-square of 3482.39 ($p<0.0001$) indicating the presence of a random effect. The residual intra class correlation coefficient (ρ) for the random intercept model was estimated to be 0.2408, which indicates that 24.1% of the unexplained variation after controlling for patient level variables could be attributed to variation between states.

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

The limited number of beneficiaries for whom ages could not be determined were not considered to be a source of any meaningful error or bias in computing the measure for each state. The majority of states had less than 1% of beneficiaries for which a date of birth was not available.

The reliability and validity of prescription claims data has been evaluated in the literature and are considered to be reliable and valid for this measure.

- Kwon et. al. found high concordance between self-report and pharmacy claims data for anti-depressant medication use (agreement 85%, kappa .069). Most discordant cases were resolved and not related to “errors” in self-report or claims data.
- Kirking et. al found in a study comparing prescription drug claims to medication use documented in medical records, that there were significantly more prescriptions documented in claims data as compared with corresponding medical records, which was even more apparent for high medication users vs. non-high users.
- A review by Lau et. al indicated that many studies have shown that pharmacy claims are more complete than medical records and are of high quality.

The hierarchical logistic regression results indicate that significant variation exists among states on the measure.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

□ **Critical data elements** (*data element validity must address ALL critical data elements*)

X **Performance measure score**

□ **Empirical validity testing**

X **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Face Validity:

PQA uses a transparent, consensus-based measure development and testing process. That process is outlined below:

- <u>Step 1</u>: PQA workgroups identify measure concepts that may be appropriate for development into fully specified performance measures. The workgroups focus on specific aspects of the medication-use system and/or specific therapeutic areas. The workgroups are open to all members of PQA and use a consensus-based approach to identify, prioritize and recommend the measure concepts that are deemed to be highly important for supporting quality improvement related to medications.
- <u>Step 2</u>: The measure concepts that are recommended for further development through a vote by the PQA workgroups are forwarded to the PQA Quality Metrics Expert Panel (QMEP) for evaluation and refinement. The QMEP reviews the measure concepts to provide an initial assessment of the key properties of performance measures (i.e., feasibility, usability and scientific validity). The measure concepts that are rated highly on these key properties will then undergo technical specification.

- Step 3: The draft measure is provided to PQA member organizations for their comments prior to preparing technical specifications for pilot testing. The QMEP reviews member comments, edits the draft measure accordingly and poses testing questions based on this all-member feedback.
- Step 4: PQA selects partners to test the draft measure. These partners are often PQA member health plans or academic institutions with expertise in quality and performance measure testing. The testing partner implements the draft technical specifications with their existing datasets and provides a report to PQA that details testing results and recommendations for modifications of the technical specifications.
- Step 5: The workgroup that developed the measure reviews the testing results and provides comment. The QMEP reviews the workgroup comments, testing results, recommendations and potential modifications and provides a final assessment of the feasibility and scientific validity of the draft performance measures.
- Step 6: Measures that are recommended by the QMEP for endorsement are posted on the PQA web site for member review, written comments are requested, and a conference call for member organizations is scheduled to address any questions. This process allows members to discuss their views on the measures in advance of the voting period.
- Step 7: PQA member organizations vote on the performance measure(s) considered for approval/endorsement.

The assessment of the *Antipsychotic Use in Children Under 5 Years Old* measure's face validity is detailed below.

The *Antipsychotic Use in Children Under 5 Years Old* measure was tested for face validity (i.e., whether it appears to measure what it intends to measure) through review by the PQA Quality Metrics Expert Panel (QMEP), PQA's full membership, and the health plans that pilot tested the measure.

The PQA QMEP is a panel that comprises individuals with expertise and experience in pharmacy, medicine, research, and clinical or other technical expertise related to quality improvement and measure development. The QMEP reviewed the measure prior to testing to ensure scientific soundness and usefulness. The QMEP also considered the age criteria, demographic criteria (foster children breakout), duration of treatment, and stepwise construction of the measure calculation. The QMEP reviewed the results of the measure testing and found the measure to be feasible. Testing results showed that rates of antipsychotic use in patients under the age of 5 varied between states and there appeared to be room for improvement. The QMEP voted unanimously to recommend that PQA members consider the measure for endorsement.

PQA membership was notified prior to the PQA Annual Meeting in May 2013, of the opportunity to consider and vote for the performance measure during the meeting. (Note: PQA membership comprises health plans, community pharmacy, long-term care pharmacies, HIT companies, PBMs, healthcare quality and standards organizations, professional and trade associations, and others. Visit http://www.pqaalliance.org/members.htm for the full list of groups and member organizations.) Members received the measure description, key points and evidence, and measure specifications. During the PQA Business meeting, the measure was reviewed and there was open discussion of the measure. Nearly all of PQA membership had a representative at the Annual Meeting and were present for the vote. Voting options included, "Agree" (indicating that the organization approved the measure), "Disagree (indicating that the organization opposed the measure) and "Abstain." The vast majority of the membership voted in favor of approving the measure.

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

N/A

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The PQA measure development process is designed to assure face validity of proposed measures; thus the measure was considered to have face validity.

The measure also was reviewed and tested by a University with access to Medicaid data. Through this testing and analysis, the measure was confirmed to have face validity. In addition, in particular, the University confirmed after testing that they were able to utilize the data to easily and accurately determine age at the time the prescriptions were dispensed, which then was used to identify the subset of patients who were under 5 years of age at the time of prescription dispensing.

____

2b3. EXCLUSIONS ANALYSIS
NA X no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

N/A

2b4.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is <u>not risk adjusted or stratified</u>, provide <u>rationale and analyses</u> to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.
2b4.3. Describe the conceptual/clinical <u>and</u> statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)
2b4.4. What were the statistical results of the analyses used to select risk factors?
2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model <u>or</u> 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 <u>2b4.9</u>

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
2b4.9. Results of Risk Stratification Analysis:
2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
2b4.11. Optional Additional Testing for Risk Adjustment (*<u>not required</u>, 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*)

____

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We hypothesized that there are differences between the states in prescribing rates for antipsychotics in children less than 5 years old. We used 2007 Medicaid data to calculate the mean, median, standard deviation, and interquartile range for the prescribing rates for 44 states and the District of Columbia. We divided the rates in quartiles, and compared the prescribing rates between the bottom quartile (75%) and top quartile (25%) with a Student's t-test. In addition, we created 4 groups based on the quartiles, and conducted a one way analysis of variance (ANOVA) to examine the differences in prescribing rates between the groups.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Table 1 reports the measures for variability for the state prescribing rates (mean, median, standard deviation and interquartile range). Table 2 reports the mean, standard deviation and p-values for the Student's t-test and ANOVA results.

Table 1. Variation in State antipsychotic prescribing rates for children under 5 years

Measure	Result (%)
Mean	0.1242
Median	0.1104
Standard Deviation	0.0767
Interquartile Range	0.0718

Table 2. Differences in State antipsychotic prescribing rates for children under 5 years

	Mean (%)	Standard Deviation	p-value (Student's t-test)	p-value (ANOVA)
Quartile 1	0.0480	0.0350	<0.0001*	<0.0001**
Quartile 2	0.0929	0.0095		
Quartile 3	0.1302	0.0119		
Quartile 4	0.2326	0.0572	<0.0001*	

* Student's t-test comparing differences in the top and bottom quartiles

** ANOVA test comparing differences between all four quartiles

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Antipsychotic prescribing rates in children under 5 in the Medicaid population are significantly different from state to state. The mean and median of the population prescribing rates differ by 11.1%, and the standard deviation is 0.07%, which indicates some variation in this population. There is a statistically significant difference in prescribing rates between the top and bottom quartile of the population, as well as between all four quartile groups (for both groups p<0.0001 at alpha=0.05).

Studies demonstrate that children who receive antipsychotic medications have a greater risk of both immediate and long-term complications, and adverse effects on health, including diabetes, metabolic, and cardiovascular issues (i.e., increases in weight, BMI, total cholesterol, triglycerides, QTc interval, heart rate, and liver function abnormalities), as well as hyperprolactemia, thyroid dysfunction, depression, agranulocytosis, and extrapyramidal syndromes.

As stated above, antipsychotic prescribing rates in children under 5 years old in the Medicaid population are significantly different from state to state, and show statistically significant differences between all four quartile groups. Important to note is that if all 45 states included in the testing had the average prescribing rate of the 75th percentile, we would reduce harm in 7,432 children.

____

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

- Only one set of specifications is provided for this measure.

Note*: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and what are the norms for the test conducted*)

____

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The only missing data analysis that could be tested was for age. Age was determined from the date of birth field in the Personal Summary enrollment files for each state. The inclusion criteria for the denominator of this measure was one (1) month of enrollment and less than age 5 at any time during the observation year. Analysis of missing data was based on the percentage of beneficiaries meeting the enrollment criteria and having a missing date of birth. Date of birth was not available for only 1.8% of beneficiaries in the 45 states for the national analysis. 25 states had less than 1.0% missing, 9 states had 1.0-1.9% missing, and 11 states had 2.0% or more missing.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Age data were not available for only 1.8% of beneficiaries in the 45 states for the national analysis. 25 states had less than 1.0% missing, 9 states had 1.0-1.9% missing, and 11 states had 2.0% or more missing.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; <u>if no empirical analysis</u>, provide rationale for the selected approach for missing data*)

The limited number of beneficiaries for whom ages could not be determined were not considered to be a source of any meaningful error or bias in computing the measure for each state. The majority of states had less than 1% of beneficiaries for which a date of birth was not available. There was no systematic pattern to missing data, therefore, no potential bias was considered to exist.