NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2806
Measure Title: ~~Pediatric~~ Adolescent Psychosis: Screening for Drugs of Abuse in the Emergency Department
Measure Steward: Seattle Children´s Research Institute
Brief Description of Measure: Percentage of children/adolescents age ~~=~~12~~5~~ to ~~=~~19 years-old seen in the emergency department with psychotic symptoms who are screened for alcohol or drugs of abuse
Developer Rationale: In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.
The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of quality measures related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. Pediatrics 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is an indicator designed to fill this key measurement gap.
The COE4CCN Mental Health Working Group (see item Ad.1 for more details on this group) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group. We found that psychosis was the third most common reason for pediatric mental health hospitalizations (Bardach et al. Pediatrics 2014). Literature reviews were then conducted separately for each of the most common conditions, and one of these reviews focused on children evaluated and treated for psychosis in the ED and inpatient settings. See Evidence form for conceptual model underlying the rationale for the measures.
Based on the literature reviews, we developed a list of draft quality measures to assess the quality of pediatric mental health care in the ED and inpatient settings, including specific measures to assess the quality of care for children presenting with psychotic symptoms. The validity and feasibility of these indicators were then evaluated by an expert panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in 5 hospitals in Washington state, Ohio, and Minnesota. This measure submission presents the results of this development and field testing work.

Numerator Statement: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests.
Denominator Statement: Patients aged ~~=5~~ 12 to ~~=~~19 years-old seen in the emergency department with psychotic symptoms.
Denominator Exclusions: No patients were excluded from the target population.

Measure Type: Process
Data Source: Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: N/A

Preliminary Analysis

The preliminary analysis was developed in response to recommendations from NQF's Consensus Task Force and measurement stakeholders as a way to enhance and streamline the measure evaluation and voting processes. The preliminary analysis, developed by NQF staff, will help to guide the Standing Committee evaluation of each measure by summarizing the measure submission and identifying topic areas for discussion. **NQF staff would like to stress that the preliminary analysis is intended to be used as a guide to facilitate the Committee's discussion and evaluation.**

Criteria 1: Importance to Measure and Report

1a. evidence

1a. Evidence. The evidence requirements for a *process* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The evidence for this process measure should demonstrate that the process of checking for drugs of abuse for a patient who presents with psychotic symptoms should improve outcomes and limit missed diagnoses, lack of treatment, and representation to care.

The developer provides the following information for this facility-level process measure:

- The developer cites a 2013 guideline from the American Academy of Child and Adolescent Psychiatry (AACAP): "*Clinical Practice Guideline Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, substance abuse, developmental disabilities, psychosocial stressors, and medical problems. [CS]*
    - There are no neuroimaging, psychological, or laboratory tests that establish a diagnosis of schizophrenia. The medical evaluation focuses on ruling out nonpsychiatric causes of psychosis and establishing baseline laboratory parameters for monitoring medication therapy. ... Toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out."
    - The recommendation carries AACAP's highest grade of clinical standard—i.e., based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials, and/or overwhelming clinical consensus).
    - The guideline does not provide citations for the recommendation, so there is no summary on the quantity, quality, and consistency of the evidence nor a grade. The recommendation's highest grade is derived from overwhemling clinical consensus.
- The developer provides no additional reviews or literature, indicating no studies were identified since AACAP published the guideline in 2013.
- Per the **NQF Algorithm for Evidence,** there is no systematic review (box 3) and no additional empirical evidence submitted (box 7). The Committee's evaluation should focus on whether the rating should be INSUFFICIENT WITH EVIDENCE EXCEPTION or INSUFFICIENT (boxes 10-->12).

Questions for the Committee

- *Are there (OR could there be) performance measures of a related health outcome, OR evidence-based clinical intermediate outcome?*
- *Is there evidence of a systematic opinion (e.g., national/international consensus recommendation) that the benefits of what is being measured outweigh potential harms)?*
- *Does the Steering Committee agree that it is OK (or beneficial) to hold providers accountable in the absence of empirical evidence of benefits to patients?*

1b. Gap in Care/Opportunity for Improvement and 1b. disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer provides the following information:

- Mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 in 2009, with psychosis the third most common mental health diagnosis (12.1%).
- Performance gap information was derived from testing the measure using data aggregated over two years from three children's hospitals and two community hospitals. Included patients were discharged from one of the

hospital EDs during the two year measurement period (January 1, 2012-December 31, 2013). The performance scores are presented below:

# of hospitals: 5
~~# of patients: 257~~
~~Mean hospital-level score (0-100 scale): 28.8~~
~~95% Confidence interval: 24.5-33.1~~
~~Min-Max: 17.8-83.3~~
# of hospitals: 5
# of patients: 209
Mean hospital-level score (0-100 scale): 30.6
95% Confidence interval: 26.0-35.2
Min-Max: 20.6-88.2

- Differences were measured in performance scores by gender, race, insurance type, and chronic disease category (measured using the Pediatric Medical Complexity Algorithm.
- Using linear regression, the developer found~~-~~ that Race/ethnicity was associated with performance. The four racial ethnic categories used in the analysis were White (53%), Hispanic (1.0%), Black (29%), and Other (13%, consisting of the following subgroups: Asian/Pacific Islander, Native American, Other, Multiracial). "Other" patients were more often tested (44.4%, n=27) than White patients (27.5%, n=111); a difference in performance of 17.0% (95% CI 3.2%-30.8%). The confidence interval and statistical testing were generated using linear regression. ~~chronic disease category was associated with performance, with patients with non-complex chronic conditions more often tested (24.6%, N=67) than children with only an acute condition (15.5%, N=55) or children with a complex chronic condition (16.9%, N=80), with a difference in performance of 9.2 (95% CI 0.1-18.2) compared to patients with acute conditions only~~.
- The developer noted no other statistically significant differences by patient socio-demographic characteristics from its testing.

Questions for the Committee

- *Is there a gap in care that warrants a national performance measure?*
- *Since no disparities were identified during testing, is the Committee aware of evidence that disparities exist in this area of healthcare?*
- *Should this measure be indicated as disparities sensitive? (NQF tags measures as disparities sensitive when performance differs by race/ethnicity [current scope, though new project may expand this definition to include other disparities [e.g., persons with disabilities])*

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence.

- No directly applicable evidence available. The guideline has an "out" in it: "when exposure to drugs of abuse cannot otherwise be ruled out" - makes it hard to know what the rate should be.
- I think that the premise is that substance abuse can co-occur with schizophrenia. That is common in the adult population, but the work-up of new onset psychosis in children (especially those in the age group in which schizophrenia is very uncommon) should look for non-psychiatric causes first and there are many classes of drugs that are not drugs of abuse that when either taken in too large doses or ingested by children can result in psychosis. Steroids, ACE inhibitors, stimulant medication etc. can do this. Presentation of psychosis in the ED in children should first rule out medical causes including ingestions or inadvertent overdoses of classes of drugs that can cause psychosis as should other brain pathology. In the ED while the behavior issues around psychosis are the same for schizophrenia and medical causes the risks of harm and death from drug effects is more urgent. This measure not only has no evidence to support it, but it fails to recognize the important medical issues that might cause this symptom. A better measure would be to look for use or ingestion of any drug that might cause psychotic symptoms, not just drugs of abuse looking for the co-occurrence of conditions. The differential

diagnosis of new onset psychosis is far wider than psychiatric disorders and ruling out medical causes with different treatments other than antipsychotics is important. I didn't find any guidelines for evaluation of psychosis in children at the ED level.

- Recommendation 3 from AACAP states that screening is indicated when “exposure to drugs of abuse cannot otherwise be ruled out”.
- The recommendation carries the highest grade of clinical standard, overwhelming consensus of best practice
- There is limited evidence to support the use of a drug/alcohol screen for patients with psychotic symptoms. The evidence is based on 2013 guideline from the American Academy of Child and Adolescent Psychiatry (AACAP): Clinical Practice Guideline Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, substance abuse, developmental disabilities, psychosocial stressors, and medical problems. The guideline does not provide citations for the recommendation, so there is no summary on the quantity, quality, and consistency of the evidence nor a grade. The recommendation’s highest grade is derived from clinical consensus."

1b. Performance Gap.

- I am surprised at the low rate of testing found by the developer. Somewhat variable (wide min-max range, but CI not so wide). This seems less than optimal - would be good to have a better understanding of why this is occurring.
- The small number of patients, unclear whether or not it includes kids that presented with psychosis, but didn't have disease, makes it difficult to say much of anything useful about this measure. The sample was too small to outline disparities as it was too small to divide into groups and be statistically significant. Also question whether or not this measure belongs in psychiatry or in emergency medicine with the focus on identifying a cause for the symptoms and treating as indicated (e.g. lupus would require different treatment than drug ingestion which is different than schizophrenia). As well schizophrenia is relatively rare in children especially younger ones.
- There is a performance gap not related to socio-demographic differences
- Not enough information available to tag as disparities sensitive.
    - Measured over a 2 year period at 3 children's hospitals and 2 community hospitals.
    - Hospital mean level score was 28.8

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

The developer provides the following information:

- This is a facility-level measure; higher score = better quality.
- The data sources are administrative claims and electronic health records and paper medical records. The developer provides an attachment for the applicable codes.
- The developer defines the numerator as: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests. The denominator is defined as: Patients 12~~5~~ to 19 years seen in the emergency department with psychotic symptoms. There are no denominator exclusions, and patients are identified from hospital administrative data.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if 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 enough to distinguish differences in performance across providers.

The developer provides the following information:

- Empirical testing for reliability was conducted at a critical data elements level and performance measure score level.
- Testing was conducted at five facilities (Seattle Children's Hospital, Cincinnati Children's Hospital, University of Minnesota Children's Hospital, Fairview Ridges Hospital (MN), and Maple Grove Hospital (MN) using 2-year retrospective data (Jan 2012-Dec 2013); N=20957 patients.
  - Critical data elements were tested using inter-rater reliability of medical record abstraction.
    - The total population sample size was N=20957
    - For this specific measure, however, the sampling N=4 patients—too few to calculate a Kappa. The developer reports, however, 100% agreement.
  - Performance measure score reliability was assessed using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise)
    - ICCs were computed using STATA SE 13.
    - The developer reports the hospital-level ICC=0.442 (95%CI 0.176-0.743); N=5 hospitals
    - The developer reports that ICCs ≥0.10 indicate that there are meaningful between-site performance differences.
- Per the **NQF Algorithm for Reliability**, empirical testing was performed at the level of the computed performance measure score and so the eligible ratings are HIGH, MODERATE, or LOW (box3-->6)

Questions for the Committee

- *Does the Committee concur with the developer's conclusion that the results demonstrate sufficient reliability so that differences in performance can be identified?*

Comment [NB1]: All these patients were within the 12-19 age range.

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

- The goal of the measure is to improve outcomes for pediatric patients admitted with psychotic symptoms, which should improve outcomes and limit missed diagnosis, lack of treatment, and representation to care.
- The numerator is: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests. The denominator is: Patients aged 125 to 19 years seen in the emergency department with psychotic symptoms. There were no denominator exclusions. Patients are identified from hospital administrative data.
- The evidence for the specifications provided by the developer centers on an AACAP recommendation that is based on "overwhelming clinical consensus."
- The specifications appear consistent with the AACAP recommendation, which notes, "Toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out."

Question for the Committee

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The developer provides the following information:

- The developer tested face validity of the performance measure score. (Note, the developer checks testing of critical data elements, but then indicates no empirical testing was done. The material describes the developer's ICD conversion process.)
    - The developer performed systematic face validity assessment (RAND-UCLA Modified Delphi) of whether panelists "would consider providers who adhere more consistently to the quality measure to be providing higher quality care," which we interpret as face validity assessment at the level of the **computed measure score** (as required by NQF).
- The panelists concluded there was face validity, although other factors were bundled with the assessment.
- Per the **NQF Algorithm for Validity**, when relying only on face validity, the eligible ratings are MODERATE OR LOW (box 4-->5).

Questions for the Committee

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- No exclusions

2b4. Risk adjustment:

- No risk adjustment or risk stratification

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The developer provides the following information:

- The developer tested the difference in performance across the five hospitals using an omnibus test for difference, and then performed individual comparisons between each hospital's performance and the mean of all other hospitals.
- The developer used used ANOVA testing for the omnibus test, and a t-test to assess for individual comparisons between each hospital and the mean of all others.
    - The developer indicates the results detect statistically and clinically meaningful differences in hospital performance.

Question for the Committee

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- The developer notes is unlikely that missing data contributes to substantial or meaningful biases of performance estimates. The two potential areas for missing data are at the level of the administrative claims and medical abstraction stage. Missing data in the medical abstraction stage are interpreted as the patient not meeting the measure specifications.
    - The developer posits it would be very unusual for a laboratory test (urine or serum) to be sent, processed, and not documented given the regulations around laboratory and quality insurance, as well as the need to be reimbursed for the testing.

- The developer concludes there is unlikely to be a substantial incidence of false negatives for the measure due to missing data or biased performance results due to differentially missing data.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. Specifications

- A sample of four is too small to make any determinations from. Even if they all measured "drugs of abuse" it doesn't mean that they were measuring the same thing or the same set of drugs. "drugs of abuse" is not definitive and therefore difficult to reproduce without further definition. Unclear if they considered anabolic steroids which can be abused, but are typically not drugs of abuse from a substance abuse standpoint.
- Reliability testing:
  - Critical data elements were tested on only 4 subjects (100% agreement)
  - Performance measure reliability at the hospital level (n-5), ICC = 0.42.

2a2. Reliability testing

- Whether the numerator is drug AND alcohol testing or drug OR alcohol testing is not clear - stated differently in different places.
- Denominator is based on ER diagnoses which seems adequate.
- Drug screens vary in terms of the drugs that are included in the panel. The measure doesn't list the particular drugs that they are referring to except to call them "drugs of abuse" and to talk about co-occurring substance abuse. It would be difficult to know if the same drugs were being measured.
- Looks at whether or not results reported, not whether or not they're used by clinicians. Why is it a composite score (i.e., partial credit if only 1 of the 2 tested) and not "all or none"?
- Clearly defined
- Data sources are administrative claims and electronic health records and paper medical records. Applicable codes are available by developer. Specifications seem appropriate:
  - numerator is: Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests.
  - denominator is defined as: Patients 12~~5~~ to 19 years seen in the emergency department with psychotic symptoms. There are no denominator exclusions, and patients are identified from hospital administrative data.
  - only concern may be with paper charts"

2b1. Validity Specifications

- There is no consideration of the "out" that is provided in the guideline (which is the only evidence supporting the measure).
- While the specifications may be consistent with the evidence, the limitation of toxicology testing to drugs of abuse and the focus on co-occurring mental illness and substance abuse in the documents belie the fact that psychosis may be exposure to a class of drugs not related to abuse and not in fact related to schizophrenia at all.
- Evidence for the specifications is based more on clinical consensus rather than scientific evidence.

2b2. Validity Testing

- No empirical validity testing done. Score from Delphi group acceptable but on the low side.
- The validity of this measure is confounded as it is measured with other factors. The sample size and number of hospitals is small also making conclusions difficult to make. As well it is unclear that this measure improved outcome, function or treatment since they were only looking for co-occurring substance use and not psychosis related to other drugs.
- It looks like 78.6% of the patients in the validation set came from 2 of the 5 hospitals. Is this a broad enough population?
- Face validity measured per developer, not a lot of information given
- Face validity is sufficient

2b3-2b7. Threats to Validity

- Not likely - agree with developer that these are very clear data elements.
- While it would be difficult to lose a lab test, it is unclear that the lab tests would all be the same across the country since toxicology screens differ between regions, hospitals, and labs. Unclear that this constitutes quality care as there are no specifics for what is being tested for, what is considered abnormal, and how the

information is being used.
- Two potential areas for missing data are at the level of the administrative claims and medical abstraction stage. Lab tests are typically documented in medical record.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer provides the following information:

- ALL data elements are in defined fields in a combination of electronic sources.
- Data are generated or collected by and used by healthcare personnel during the provision of care.

Questions for the Committee

- *Do you concur that the required data elements are routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use?*

Committee pre-evaluation comments
Criteria 3: Feasibility

- Feasible
- Data already collected , feasible to extract from electronic sources
- Electronic records and claims should include such testing however they are unlikely to include the details of the testing (tox screens vary and so may not be measuring the same things).
- Data elements are defined fields in EMR and collected during treatment.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

- This measure is not in use. It has not been implemented as the development, validation, and testing were just recently completed.
- Planned use include: Quality Improvement with Benchmarking (external benchmarking to multiple organizations) and quality Improvement (Internal to the specific organization)
- There were no unintended consequences identified during testing.

Questions for the Committee*:*

- *The developer indicates use for benchmarking and quality improvement. NQF endorsement focuses on primarily accountability, and then appropriateness for quality improvement. Is this measure appropriate for accountability purposes?*
- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Committee pre-evaluation comments
Criteria 4: Usability and Use

- No unintended consequences
- This measure is incomplete for the appropriate emergent evaluation of psychosis as it excludes looking for classes of drugs that are not drugs of abuse. It is important to look for co-occurring substance abuse (or psychosis related to drugs of abuse), but that is only part of the equation. Using a measure that doesn't include all of the possibilities gives the impression that this is all that is necessary to provide quality care.
- Not in use yet, would be good for quality improvement/benchmarking.

Criterion 5: Related and Competing Measures
• No related and competing measures

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: ~~**Pediatric**~~ **Adolescent Psychosis**: Screening for Drugs of Abuse in the Emergency Department

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:

Date of Submission: 9/30/2015

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *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 10 pages (*incudes 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 evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the

strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ **Process**: Screening for drugs of abuse for ~~pediatric~~ adolescent patients who present to the Emergency Department with symptoms of psychosis.

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

N/A

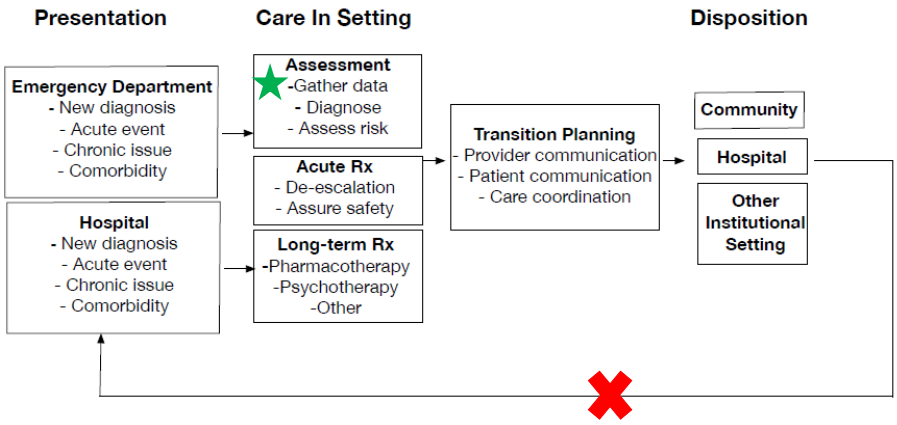
1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

N/A

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.



This diagram depicts the relationship between the care process of interest, marked with a green star, and the target outcomes to prevent (rehospitalizations and re-presentations to the ED), marked with the red X. The proposed measure focuses on whether one element of "Gather data" (Assessment box) was performed. If the process of checking for drugs of abuse for a patient who presents with psychotic symptoms is not performed, this may lead to a missed diagnosis, lack of treatment, and representation to care.

Summary: Overall, there is not extensive empirical literature supporting this process measure, but the benefits likely far outweigh the risks.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

McClellan J, Werry J, Bernet W, Arnold V, Beitchman J, Benson RS, Bukstein O, Kinlan J, Rue D, Shaw J, Kroeger K: Practice parameter for the assessment and treatment of children and adolescents with schizophrenia, Journal of the American Academy of Child and Adolescent Psychiatry 2013, Volume 52, Issue 9, Pages 976–990

http://www.jaacap.com/article/S0890-8567(13)00112-3/fulltext

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, ***substance abuse***, developmental disabilities, psychosocial stressors, and medical problems. [CS]

Youth with suspected schizophrenia require a thorough psychiatric and medical evaluation, including the assessment for common comorbid conditions, such as substance abuse or cognitive delays. When present, active psychotic symptoms are generally prioritized as the main target for treatment. Comorbid conditions, such as substance abuse, may respond better to treatment once acute symptoms of schizophrenia are stabilized. However, any life-threatening symptoms, such as suicidal behavior or severe aggressive behaviors, must be prioritized in the treatment plan.

There are no neuroimaging, psychological, or laboratory tests that establish a diagnosis of schizophrenia. **The medical evaluation focuses on ruling out nonpsychiatric causes of psychosis** and establishing baseline laboratory parameters for monitoring medication therapy. More extensive evaluation is indicated for atypical presentations, such as a gross deterioration in cognitive and motor abilities, focal neurologic symptoms, or delirium.

Assessments are obtained based on specific medical indications, e.g., neuroimaging studies when neurologic symptoms are present or an electroencephalogram for a clinical history suggestive of seizures**. Toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out**. Genetic testing is indicated if there are associated dysmorphic or syndromic features. Similarly, tests to rule out specific syndromes or diseases (e.g., amino acid screens for inborn errors of metabolism, ceruloplasmin for Wilson disease, porphobilinogen for acute intermittent porphyria) are indicated for clinical presentations suggestive of the specific syndrome in question. Broad screening for rare medical conditions is not likely to be informative in individuals with psychosis who do not present with other neurologic or medical concerns.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

The AACAP guidelines granted this their highest grading:

•Clinical Standard [CS] is applied to recommendations that are based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials) and/or overwhelming clinical consensus

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

•Clinical Guideline [CG] is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized controlled trials, cohort studies, case-control studies) and/or strong clinical consensus

•Clinical Option [OP] is applied to recommendations that are based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion, but lack strong empirical evidence and/or strong clinical consensus

•Not Endorsed [NE] is applied to practices that are known to be ineffective or contraindicated

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Recommendation 3. Youth with suspected schizophrenia should be carefully evaluated for other pertinent clinical conditions and/or associated problems, including suicidality, comorbid disorders, *substance abuse*, developmental disabilities, psychosocial stressors, and medical problems. [CS]

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The AACAP guideline does not provide citations for the Recommendation and so there is no grade assigned for the quality of the quoted evidence to support the Recommendation. The specific endorsement of drugs of abuse screening within Recommendation 3 is therefore not supported with citations of evidence. Nevertheless, the guidelines granted the Recommendation overall the highest grading of Clinical Standard [CS] (defined below). Thus, this recommendation is bolstered by overwhelming clinical consensus.

"Clinical Standard [CS] is applied to recommendations that are based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials) and/or overwhelming clinical consensus"

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

•Clinical Guideline [CG] is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized controlled trials, cohort studies, case-control studies) and/or strong clinical consensus

•Clinical Option [OP] is applied to recommendations that are based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion, but lack strong empirical evidence and/or strong clinical consensus

•Not Endorsed [NE] is applied to practices that are known to be ineffective or contraindicated

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

NA

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

NA

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

NA

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

NA

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

NA

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

No studies providing new evidence to support this quality measure were identified since the publishing of the AACAP guideline in 2013.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
P2_Screen_for_Tox_evidence_attachment_2015_09_30_FOR_SUBMISSION.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- Disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
In March 2011, the Centers for Medicare and Medicaid Services (CMS) and the Agency for Healthcare Research and Quality (AHRQ) partnered to fund seven Centers of Excellence on Quality of Care Measures for Children (COEs). These Centers constitute the Pediatric Quality Measures Program (PQMP) mandated by the Child Health Insurance Program Reauthorization Act (CHIPRA) legislation passed in January of 2009. The charge to the seven COEs is to develop new quality of care measures and/or enhance existing measures for children's healthcare across the age spectrum.
The Center of Excellence on Quality of Care Measures for Children with Complex Needs (COE4CCN), in response to a charge from CMS and AHRQ, developed a set of quality measures related to the management of children and adolescents with mental health problems presenting to the emergency department (ED) and inpatient settings. CMS and AHRQ's choice of mental health as a focus for measurement reflects the dearth of measures in pediatric mental health (Zima et al. Pediatrics 2013) and the importance of optimizing treatment for these illnesses. The proposed measure is an indicator designed to fill this key measurement gap.
The COE4CCN Mental Health Working Group (see item Ad.1 for more details on this group) first conducted secondary analyses of national and state-based data to identify the most common mental health diagnoses resulting in hospitalization in the pediatric age group. We found that psychosis was the third most common reason for pediatric mental health hospitalizations (Bardach et al. Pediatrics 2014). Literature reviews were then conducted separately for each of the most common conditions, and one of these reviews focused on children evaluated and treated for psychosis in the ED and inpatient settings. See Evidence form for conceptual model underlying the rationale for the measures.
Based on the literature reviews, we developed a list of draft quality measures to assess the quality of pediatric mental health care in the ED and inpatient settings, including specific measures to assess the quality of care for children presenting with psychotic symptoms. The validity and feasibility of these indicators were then evaluated by an expert panel (see Item Ad.1) using the RAND-University of California, Los Angeles (UCLA) modified Delphi method (see Testing form for description of Delphi process used), and subsequently field tested in 5 hospitals in Washington state, Ohio, and Minnesota. This measure submission presents the results of this development and field testing work.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In a field test of this quality measure, performed as part of the funded development work, we measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital and from two community hospitals in Minnesota, Fairview Ridges Hospital and Maple Grove Hospital. Included patients were discharged from one of the hospital EDs during the two year measurement period (January 1, 2012-December 31, 2013). The performance scores are presented below.

# of hospitals: 5
# of patients: 209
Mean hospital-level score (0-100 scale): 30.6
95% Confidence interval: 26.0-35.2
Min-Max: 20.6-88.2

See Testing form, item 2b.5.2a for data on individual hospital performance.

~~# of hospitals: 5~~
~~# of patients: 257~~
~~Mean hospital-level score (0-100 scale): 28.8~~
~~95% Confidence interval: 24.5-33.1~~
~~Min-Max: 17.8-83.3~~

~~See Testing form, item 2b.5.2a for data on individual hospital performance.~~

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
In the field testing described above, we measured differences in performance scores by gender, race, insurance type, and chronic disease category (measured using the Pediatric Medical Complexity Algorithm—Simon et al. Pediatrics 2015). Race/ethnicity was associated with performance. The four racial ethnic categories used in the analysis were White (53%), Hispanic (1.0%), Black (29%), and Other (13%, consisting of the following subgroups: Asian/Pacific Islander, Native American, Other, Multiracial). "Other" patients were more often tested (44.4%, n=27) than White patients (27.5%, n=111); a difference in performance of 17.0% (95% CI 3.2%-30.8%). The confidence interval and statistical testing were generated using linear regression.~~In the field testing described above, we measured differences in performance scores by gender, race, insurance type, and chronic disease category (measured using the Pediatric Medical Complexity Algorithm—Simon et al. Pediatrics 2015). Chronic disease category was associated with performance, with patients with non-complex chronic conditions more often tested (24.6%, n=67) than children with only an acute condition (15.5%, n=55) or children with a complex chronic condition (16.9%, n=80), with a difference in performance of 9.2 (95% CI 0.1-18.2) compared to patients with acute conditions only. The confidence interval and statistical testing were generated using linear regression.~~

There were no other statistically significant differences by patient socio-demographic characteristics in our testing. Please see Testing form, item 2b.5.2b for data.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Psychosis in pediatric patients is a high priority aspect of healthcare, with substantial inpatient utilization and high severity of illness, in addition to a number of associated costs to the healthcare system and to patients and families. Mental health hospitalizations for pediatrics represented 9.1% of all hospitalizations for children ages >2 in 2009, with psychosis the third most common mental health diagnosis (12.1%), after depression (44.1%) and bipolar disorder (18.1%).1 A significant increase in the diagnosis of psychotic

disorders from 8.3 to 12.0 percent of hospital discharges was found in a national survey of inpatient mental health services for children and adolescents from 1999 to 2000.2 Specific predictors of poor long term outcomes include more than two inpatient-treated episodes of schizophrenia3 and a longer duration of first inpatient treatment.3 Lay et al.3 found that 12 years after their initial diagnoses of schizophrenia only 17% of adolescents had not been readmitted for further inpatient treatment, and there was a median of 4 subsequent inpatient-treated episodes. Similarly, Fleischhaker et al.4 found an average of 3 readmissions for 40% of patients in a 10-year follow-up for adolescent-onset schizophrenia.

Children and adolescents with a diagnosis of a psychotic disorder face a number of challenges medically, socially, and developmentally. Several studies found a high risk of educational and/or occupational impairment for patients with early-onset schizophrenia.3,4

A number of costs have been associated with early-onset psychosis for the medical system as well as the patient and family. Length of stay for inpatients with psychosis has been found to typically be longer than for other mental health diagnoses.5 In addition, in a comparison of mental health versus non-mental health ED visits from 2001-2008, patients with a mental health diagnosis had fewer referrals to outpatient care5 and a higher number of inpatient admissions.5 Long-term studies of patients with early-onset psychosis have found that as adults, most were financially dependent on family or receiving public assistance.3,4

In the proposed measure, we specifically focus on the issue of comorbid substance abuse in this population. The American Academy of Child and Adolescent Psychiatry (AACAP) recommends that youth with suspected schizophrenia require a thorough psychiatric and medical evaluation, including the assessment for common comorbid conditions, such as substance abuse or cognitive delays,6 specifying that toxicology screens are indicated for acute onset or exacerbations of psychosis when exposure to drugs of abuse cannot otherwise be ruled out.6 Comorbid substance abuse is common in patients with psychosis7-9 and can lead to decreased access of psychiatric services,10,11 while also leading to potentially avoidable healthcare utilization.6,11 Accurately diagnosing co-morbid substance abuse, or accurately diagnosing substance abuse presenting with psychotic symptoms, is an essential first step to appropriate management, referral, and obtaining access to services to address the substance abuse.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Bardach NS, Coker TR, Zima BT, et al. Common and costly hospitalizations for pediatric mental health disorders. Pediatrics. 2014;133(4):602-609.
2. Case BG, Olfson M, Marcus SC, Siegel C. Trends in the inpatient mental health treatment of children and adolescents in US community hospitals between 1990 and 2000. Arch Gen Psychiatry. 2007;64(1):89-96.
3. Lay B, Blanz B, Hartmann M, Schmidt MH. The psychosocial outcome of adolescent-onset schizophrenia: a 12-year followup. Schizophr Bull. 2000;26(4):801-816.
4. Fleischhaker C, Schulz E, Tepper K, Martin M, Hennighausen K, Remschmidt H. Long-Term Course of Adolescent Schizophrenia. Schizophrenia Bulletin. 2005;31(3):769-780.
5. Case SD, Case BG, Olfson M, Linakis JG, Laska EM. Length of stay of pediatric mental health emergency department visits in the United States. J Am Acad Child Adolesc Psychiatry. 2011;50(11):1110-1119.
6. McClellan J, Stock S. Practice parameter for the assessment and treatment of children and adolescents with schizophrenia. Journal of the American Academy of Child & Adolescent Psychiatry. 2013;52(9):976-990.
7. Hsiao R, McClellan J. Substance abuse in early onset psychotic disorders. Journal of Dual Diagnosis. 2008;4(1):87-99.
8. Cannon TD, Cadenhead K, Cornblatt B, et al. Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. Arch Gen Psychiatry. 2008;65(1):28-37.
9. Regier DA, Farmer ME, Rae DS, et al. Comorbidity of mental disorders with alcohol and other drug abuse: Results from the epidemiologic catchment area (eca) study. JAMA. 1990;264(19):2511-2518.
10. Dyck DG, Hendryx MS, Short RA, Voss WD, McFarlane WR. Service use among patients with schizophrenia in psychoeducational multiple-family group treatment. Psychiatr Serv. 2002;53(6):749-754.
11. Schooler NR, Keith SJ, Severe JB, et al. Relapse and rehospitalization during maintenance treatment of schizophrenia. The effects of dose reduction and family treatment. Arch Gen Psychiatry. 1997;54(5):453-463.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when

implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Behavioral Health, Behavioral Health : Alcohol, Substance Use/Abuse, Behavioral Health : Screening, Behavioral Health : Serious Mental Illness, Mental Health, Mental Health : Alcohol, Substance Use/Abuse, Mental Health : Serious Mental Illness

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Measure specifications can be found at the following URL under the heading: "Mental Health Measures":
http://www.seattlechildrens.org/research/child-health-behavior-and-development/mangione-smith-lab/measurement-tools/

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PSYCHOSIS_ICD9_and_ICD10_Codes_for_Denominator_Identification_SUBMITTED-635803493103736421.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Eligible patients with documentation of drug and alcohol screening using urine drug or serum alcohol tests.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
24 month period of data, retrospectively collected. We propose using 24 months due to the low prevalence of the condition. This is the period used in the field testing of the measure.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients passing the quality measure are identified during medical record abstraction using the guidelines below. The item numbers match the "Medical Records Abstraction Tool Guidelines" under "Mental Health Measures"provided on the website in S.1. This language is also in the "Medical Records Electronic Abstraction and Scoring Tool" on the website.
11. Urine Drug Screening /Serum Alcohol Screening – [Module: Psychosis, ED care] This item applies to children and adolescents presenting with psychotic symptoms who were admitted to the marker ED. Indicate if the patient had a urine drug screen and/or serum alcohol screen while in the ED. The alcohol test will be a separate test from the drug tests. The drug test must be comprehensive in that it tests for multiple types of illicit drugs. Do NOT give credit for tests that include results of just a single drug. Drug screens commonly include tests for benzodiazepines, barbiturates, methamphetamine, cocaine, methadone, opiates, tetrahydrocannabinol, etc.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients aged 12~~=5~~ to ~~=~~19 years-old seen in the emergency department with psychotic symptoms.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Cases are identified from hospital administrative data.

Patients aged ~~=5 =~~12-19 years-old

Patients have at least one of the following ICD9 codes for psychosis, as a primary or secondary diagnosis: 291.3, 291.5, 292.11, 292.12, 293.81, 293.82, 295.30, 295.31, 295.32, 295.33, 295.34, 295.40, 295.41, 295.42, 294.43, 295.44, 295.70, 295.71, 295.72, 295.73, 295.74, 295.90, 295.91, 295.92, 295.93, 295.94, 296.24, 296.44, 297.1, 297.2, 297.3, 298.X
These codes were chosen by Members of the COE4CCN Mental Health Working Group (see Ad.1) co-chaired by Psychiatric Health Services Researchers Drs. Michael Murphy and Bonnie Zima.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
No patients were excluded from the target population.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*

Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Step 1. Identify eligible population at hospital using administrative data. N=total population
Step 2. Assess patient chart for indicator status. Pass (A=1) if documentation present of urine drug testing or both urine drug testing and serum alcohol testing. Pass (B=1) if documentation present of serum alcohol testing or both urine drug testing and serum alcohol testing.
Step 3. Calculate Patient score= 100*(A+B)/2. Results=0, 50, 100
Step 4. Calculate hospital score=Sum(Patient score)/N

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. Given the low prevalence of the condition, the measured group is the entire population of eligible patients.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
There are two potential areas for missing data: at the level of the administrative claims, which are used for sampling patients, and during medical abstraction.

Administrative Claims
There are two data fields used to identify eligible patients, the diagnosis fields and the patient age. If either is missing the case is deleted.

Medical abstraction
Missing data in the medical abstraction stage is interpreted as the patient not meeting the metric.
Please see item 2b7.1 in the testing form for additional discussion of the handling of missing data.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Electronic Clinical Data : Electronic Health Record, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data collection tool is publicly available on the website in S.1. and also attached in the Appendix materials.
Title: "Medical Record Measure Electronic Abstraction and Scoring Tool" under "Mental Health Measures"

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)* Emergency Medical Services/Ambulance, Hospital/Acute Care Facility If other:
S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)* N/A
2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** P2_Testing_for_Tox_Testing_Attachment_2015_10_13_SUBMITTED.docx

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

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: ~~**Pediatric**~~ **Adolescent Psychosis**: Screening for Drugs of Abuse in the Emergency Department
Date of Submission: 9/30/2015
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of 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.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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 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]

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 (including clinical and sociodemographic factors) that influence the measured outcome 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. 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.

As described in the submission form, the validity and feasibility of the COE4CCN pediatric mental health measures were evaluated by an expert panel using the RAND-University of California, Los Angeles (UCLA) modified Delphi method.[1]

Detailed measure specifications were developed for the endorsed pediatric mental health quality measures. These specifications were then used to develop an electronic excel macro data collection tool for use with medical records data. The tool has automated scoring capability and is available on the website listed in item S.1. Abstraction and scoring guidelines are provided as an appendix to this submission.

Field Testing of the Delphi Panel Endorsed Pediatric Mental Health Quality Measures

Three tertiary care children's hospitals and two community hospitals participated in the field test of the emergency department (ED) ***Pediatric Psychosis*** Mental Health quality measures. For each hospital, two research nurses were trained to use the medical record abstraction tool and the companion abstraction tool guidelines. For training purposes, the nurses abstracted several sample charts targeting psychosis. Their abstractions were compared to gold-standard abstractions previously completed by the developer of the measure specifications. Abstractors were considered fully trained when they could reliably abstract the gold-standard medical records.

Case Selection

Cases for the field test were selected using International Classification of Diseases 9th Revision Clinical Modification (ICD-9) codes for psychosis from administrative databases from each hospital for discharges occurring between January 1st,2012 and December 31st, 2013 (see Appendix for a list of ICD-9 codes used to select cases for abstraction).

The final sample goal for psychosis was a total of 100 cases selected from the two larger hospitals and 35 from the three smaller hospitals, with 25% replacement cases in order to have adequate sample after patients were excluded during the medical record abstraction phase. Because of limited sample sizes at each hospital for psychosis, all eligible patients were included in the final sample. See **Table 2b5.1** for sample sizes in each hospital.

Medical Record Abstractions

For each hospital, two trained nurse abstractors were each assigned half of the case sample for psychosis. Data for each case were entered by the nurses into the electronic Pediatric Mental Health abstraction tool and both the raw data and auto-generated measure scores were uploaded to a central research database for further analysis.

At the two larger tertiary care hospitals, each nurse abstracted Pediatric Psychosis measures from 14 additional charts that were randomly selected from the other nurse's sample to facilitate assessment of inter-rater reliability (see inter-rater reliability testing results in **2a2.3** below). The 14 charts were among a total of 60 (10% sample) pulled for inter-rater reliability testing of quality measures we developed and tested across three different mental health diagnoses (psychosis, danger to self/suicidality, and substance abuse).

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**

☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

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

Two existing administrative datasets were used to sample patients using the ICD9 codes.

The Pediatric Health Information System (PHIS) database was used to sample the medical records from two of the children's hospitals. This is a comparative pediatric database, and includes clinical and resource utilization data for inpatient, ambulatory surgery, emergency department and observation unit patient encounters for 45 children's hospitals. (More information about PHIS is available at: https://www.childrenshospitals.org/Programs-and-Services/Data-Analytics-and-Research/Pediatric-Health-Information-System)

The hospital administrative discharge databases were used to sample the medical records from the other hospitals.

1.3. What are the dates of the data used in testing? January 1, 2012-December 31st, 2013

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

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

Three tertiary care children's hospitals and two community hospitals were included in the field test, from Washington state, Ohio, and Minnesota. The children's hospitals were: Seattle Children's Hospital, Cincinnati

Children's Hospital, and University of Minnesota Children's Hospital; the two community hospitals were in Minnesota: Fairview Ridges Hospital and Maple Grove Hospital.
These hospitals were selected as they are all member organizations of the COE4CCN multi-stakeholder consortium of organizations that took part in the Center's measure development activities.

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*)

Table 1.6 Testing: Sociodemographic Characteristics of Patients Eligible for Measurement with Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department **(N=2095~~7~~)**

	N	%
Child gender		
Male	124	59
Female	80	38
Missing	5	2
Child race/ethnicity		
Hispanic	3	1
White	111	53
Black	60	29
Other*	27	13
Missing	8	4
Insurance type		
Public	110	53
Private	87	42
Uninsured	7	3
Missing	5	2
PMCA category**		
Non-chronic condition	44	26
Non-complex chronic condition	53	31
Complex chronic condition	73	43

*"Other" includes Asian/Pacific Islander, Native American, Other, and Multiracial
** PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

Field Code Changed

	~~N~~	~~%~~
~~Child gender~~		
~~Male~~	~~150~~	~~58~~

	N	%
Female	98	38
Missing	9	4
Child race/ethnicity		
Hispanic	3	1
White	134	52
Black	76	30
Other	32	12
Missing	12	5
Insurance type		
Public	133	52
Private	106	41
Uninsured	9	4
Missing	9	4
PMCA category*		
Non-chronic condition	55	27
Non-complex chronic condition	67	33
Complex chronic condition	80	40

* PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015).[2] Available only at 2 of the 3 participating hospitals.

Field Code Changed

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.

NA

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

To measure patient-level sociodemongraphic variables, we used patient gender, race, ethnicity, insurance type, and chronic disease status. These variables were derived from the administrative claims data from each participating hospital. Chronic disease status was captured using the Pediatric Medical Complexity Algorithm (PMCA), which categorizes pediatric inpatients using diagnostic ICD9 codes as having an acute medical condition only (non-chronic condition), a non-complex chronic condition, or a complex chronic condition.[2] Retrospective claims data needed to run PMCA were only available from 2 of the 5 field test hospitals.

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*)

Critical data elements used in the measure were tested for inter-rater reliability of medical record abstraction. Reliability was measured using the prevalence adjusted bias adjusted kappa (PABAK) statistic for patient eligibility for measurement, and for the patient score for the quality measure. Kappa is a statistic that captures the proportion of agreement beyond that expected by chance, that is, the *achieved* beyond-chance agreement as a portion of the *possible* beyond-chance agreement.[3] PABAK is a measure of inter-rater reliability that adjusts the magnitude of the kappa statistic to take account of the influences of high or low prevalence and of inter-rater differences in assessment of prevalence. The PABAK statistic adjusts for high or low prevalence and is what we used in our calculations of inter-rater reliability.

Performance measure score was assessed for reliability across performance sites using the intra-class correlation coefficient (ICC). The ICC assesses the ratio of between site variation and within site variation on performance. Higher ICC implies that the between site variation (signal) is higher than the within site variation (noise). ICCs were computed using STATA SE 13.

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*)

There are two stages of medical record abstraction for which we tested inter-rater reliability for all Pediatric Mental Health Measures: patient eligibility for the measure; and patient score for the quality measure. For this measure, because there were no medical record exclusions, we did not measure patient eligibility kappas, since there were no abstractions for that stage.

The specific measure addressed in this submission was one of 6 psychosis measures included in the field test as part of the broader COE4CCN Pediatric Mental Health Measures in the Hospital Setting Project.

Across all 6 psychosis measures tested in the field, 120 records were sampled and abstracted by both nurse abstractors.

Comment [NB2]: All records were within the age range 12-19.

Kappa for patient score for all 6 psychosis measures (n=98 eligible patient charts):
0.62.
PABAK for patient score for all 6 psychosis measures (n=98 eligible patient charts):
0.71.

For the specific submitted measure, only a very small subset (n=4) of the randomly sampled charts were eligible. There were too few patients eligible for this measure to calculate kappa. Instead, we present the percent agreement.

Comment [NB3]: All records were within the 12-19 age range

Percent agreement for patient scores on the quality measure under consideration:
100%

Performance measure score:
We performed ICC testing for performance variation at the level of the hospital, since that is the intended level of measurement. However, despite adequate sample size at the patient level within each site (see Table 2b5.1 below), the number of higher level clusters in our field test is limited to the 5 participating hospitals. Future measurement across a larger number of participating hospitals will give more generalizable estimations of ICC for this measure.
Hospital-level ICC=0.44 (95%CI 0.17-0.74). N=5 hospitals
~~Hospital-level ICC=0.42 (95%CI 0.16-0.73). N=5 hospitals~~

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?*)

Critical data elements: Interpretation of Kappas is generally cited as follows[3,4]: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.
Hence, inter-rater reliability for psychosis measures was substantial. For the specific submitted measure, percent agreement was perfect.

Performance measure score: Hospital level ICC based on the five hospitals is relatively high. ICCs ≥0.10 are considered relatively high.[5] Hence, the ICCs indicate that there are meaningful between-site performance differences.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**
☐ **Empirical validity testing**
☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

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

CRITICAL DATA ELEMENTS
ICD10 CONVERSION (no testing performed)

1. Statement of intent for the selection of ICD-10 codes:
    a. **The goal is to convert this measure to a new code set, fully consistent with the intent of the original measure.**
2. Excel spreadsheet with original ICD-9 codes from the Field test and the ICD9-ICD10 conversion table is attached at S2.b
3. Description of the process used to identify ICD-10 codes, including:
    a. Experts who assisted in the process:
        i. Bonnie Zima (co-chair Mental Health Working Group, see Ad.1)

ii. Michael Murphy (co-chair Mental Health Working Group, see Ad.1)

b. Name of the tool used to identify/map to ICD-10 codes:

i. Transformation was based on the Centers for Medicaid and Medicare Services Gems tool.

c. Stakeholder input was obtained from the COE4CCN Mental Health Multi-stakeholder Working Group. See below.

Psychosis ICD9 to ICD10 Conversion: Stakeholder Comments

A) Researcher and practitioner stakeholder #1:

"Psychosis - F44.89 - I usually think of dissociative disorders and conversion as not being delusional or psychotic. They are more loss of function than hallucinations, etc. So, I am not sure that this code belongs."

Response: consultation with stakeholder #3 and then deleted this code.

B) Researcher and practitioner stakeholder #2:

"I read all the new ICD 10 dx for both psychosis and substance abuse and they all seemed appropriate. They also all seemed to correspond pretty well to their ICD 9 antecedents. I am signing off on these lists. I think that the codes make sense."

Response: none needed

C) Researcher and practitioner stakeholder #3:

"re: Psychosis - F44.89, agree with [stakeholder #1] re: conversion is a somatoform disorder. Would delete."

"re: Psychosis - F44.89, I've honestly never heard of the dx "reactive confusion" and it's not in either the DSM 5 or DSM IVR. Thus I agree with [stakeholder #1]. I also wonder whether during this exercise we are getting caught up with a more historical shift within the DSM to align with the ICD…."

Response: Deleted F44.89

D) State Medicaid office stakeholder #4:

"The mental health folks in my agency are ahead of the rest of us as they have created crosswalks that make sense for our programs. Basically the codes are being based off of the DSM-5. The DSM-5 diagnoses lists both ICD-9 and ICD-10 codes with the diagnoses."

Response: Because we went through the DSM for psychosis and chose specific ICD9s for the field testing, and there is a consistent 1:1 match with ICD9 and ICD10, we decided to keep the crosswalk for ICD9-ICD10 for psychosis.

PERFORMANCE MEASURE SCORE

EMPIRICAL VALIDITY TESTING

We did not validate this measure empirically against another measure or health outcome, due to consensus of the COE4CCN Mental Health Working Group that this is a measure of technical quality and is only one of many factors expected to ultimately influence outcomes. This measure focuses on accurate diagnosis and assessment of comorbidities which should result in more appropriate treatment and ultimately lead to beneficial changes in utilization or other directly measurable effects on health outcomes. That said, by itself, the measure was judged to be too narrow and distal from such outcomes to hypothesize a direct effect that might be tested.

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method

The content validity of the group of quality measures developed in the COE4CCN Pediatric Mental Health measures effort, which included the psychosis measure proposed, was established using the RAND-UCLA Modified Delphi Method. The process began with the nomination of 10 individuals by 8 stakeholder organizations including the American Academy of Child and Adolescent Psychiatry, the AAP Committee on Pediatric Emergency Medicine, the AAP Task Force on Mental Health, the Medicaid Medical Directors Learning Network, the AAP Section on Hospitalist Medicine, Family Voices, the Society for Adolescent Medicine, and the Substance Abuse and Mental Health Services Administration. Nine of the nominees agreed to be members of our multi-stakeholder Delphi panel. All panelists were people deemed by the nominating organizations to have substantial expertise and/or experience related to child mental health (see Ad.1 for a list of panel members). The panel read the psychosis literature review written by project staff and reviewed and scored each proposed quality measure on validity. This method is a well-established, structured approach to measure evaluation that involves two rounds of independent panel member scoring, with group discussion in between.[1] After reviewing literature review and draft psychosis quality measures, panel members were asked to rate each measure's validity on a scale from 1 (low) to 9 (high). Validity was assessed by considering whether there was adequate scientific evidence or expert consensus to support its link to better outcomes; whether there would be health benefits associated with receiving measure-specified care; whether they would consider providers who adhere more consistently to the quality measure to be providing higher quality care; and whether adherence to the measure is under the control of health care providers and/or systems. The Delphi method has been found to be reliable and to have content, construct and predictive validity.[6-10] For a quality measure or measure component to move to the next stage of measure development, it had to have a median validity score $\geq$ 7 (1-9 scale) and be scored without disagreement based on the mean absolute deviation from the median after the second round of scoring. This process ensures that only measures widely judged to be valid moved forward into measure specification. See **Table 2b.2.3** for Delphi panel scores on the measure for this submission.

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

CRITICAL DATA ELEMENTS

ICD10 CONVERSION (no testing performed)

PERFORMANCE MEASURE SCORE

SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—The RAND-UCLA Modified Delphi Method

The scores for this measure from the 9 members of the panel after round 2 of Delphi scoring (scoring done after discussions at the in-person meeting) are presented in the Table below.

Table 2b.2.3 Testing. Delphi panel: Pediatric Psychosis: **Screening for Drugs of Abuse in the Emergency Department**

	Median score	Mean absolute deviation from median	Agreement status*
Drug Screening (Urine)			
Validity	8.0	0.8	Agree
Feasibility	9.0	0.4	Agree
Alcohol screening (serum)			

Validity	7.0	1.3	Agree
Feasibility	9.0	0.4	Agree

*This is a statistical assessment of whether panelists agreed (A), disagreed (D), or if status was indeterminate (I)

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?*)

PERFORMANCE MEASURE SCORE
SYSTEMATIC ASSESSMENT OF CONTENT VALIDITY—DELPHI PANEL
The results from the Delphi panel show strong content validity for this measure, with median validity scores ≥7 (out of 9) following the Delphi panel.

2b3. EXCLUSIONS ANALYSIS
NA ☒ 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.

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

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

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 (*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*)

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)*

As noted in the Submission Item 1b, we performed a field test of the quality measure under consideration. We measured performance using data aggregated over two years from three children's hospitals, Seattle Children's Hospital, Cincinnati Children's Hospital, and University of Minnesota Children's Hospital and from two community hospitals in Minnesota, Fairview Ridges Hospital and Maple Grove Hospital. Included patients were discharged from one of the hospitals over the two year period (January 1, 2012-December 31, 2013). The performance scores are presented below in Tables 2b5.2a (performance variation across hospitals) and 2b5.2b (performance variation across socio-

demographic characteristics). We tested the difference in performance across the hospitals using an omnibus test for difference, and then performing individual comparisons between each hospital's performance and the mean of all other hospitals. We used ANOVA testing (4df) for the omnibus test, and a t-test to assess for individual comparisons between each hospital and the mean of all others.

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 2b5.2a. Performance Scores for Adolescent Psychosis: Screening for Drugs of Abuse in the Emergency Department

	Eligible patients	Hospital-level Score, Mean (95% CI)	P-value for omnibus test*	Difference from mean of all others	P-value for difference from overall mean**
Hospitals overall	209	30.6 (26.0-35.2)	<0.0001	--	--
Hospital A	34	26.5 (15.7-37.2)	--	-5.0	0.44
Hospital B	136	20.6 (16.4-24.8)	--	-28.7	<0.0001
Hospital C	17	88.2 (73.8-102.7)	--	62.7	<0.0001
Hospital D	13	61.5 (36.4-86.7)	--	33.0	0.0006
Hospital E	9	44.4 (21.3-67.5)	--	14.4	0.21

Comment [QB4]: Updated (includes age 12-19 only)

~~**Table 2b5.2a. Performance Scores for Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department**~~

	~~Eligible patients~~	~~Hospital-level Score, Mean (95% CI)~~	~~P-value for omnibus test*~~	~~Difference from mean of all others~~	~~P-value for difference from overall mean**~~
~~Hospitals overall~~	~~257~~	~~28.8 (24.5-33.1)~~	~~<0.0001~~	~~—~~	~~—~~
~~Hospital A~~	~~36~~	~~25.0 (14.7-35.3)~~	~~—~~	~~-4.4~~	~~0.48~~
~~Hospital B~~	~~166~~	~~17.8 (14.1-21.4)~~	~~—~~	~~-31.1~~	~~<0.0001~~
~~Hospital C~~	~~18~~	~~83.3 (66.3-100.4)~~	~~—~~	~~58.6~~	~~<0.0001~~
~~Hospital D~~	~~22~~	~~65.9 (47.3-84.5)~~	~~—~~	~~40.6~~	~~<0.0001~~
~~Hospital E~~	~~15~~	~~40.0 (18.6-61.4)~~	~~—~~	~~11.9~~	~~0.20~~

*Testing performed using ANOVA (4df)
**Testing performed using t-test

Table 2b5.2b. Socio-Demographic Group Scores for Adolescent Psychosis: Screening for Drugs of Abuse in the Emergency Department

Comment [QB5]: Updated (includes age 12-19 only)

	N	Mean	SD	Difference	LCL	UCL

Child gender						
Female (ref)	80	26.9	29.7			
Male	124	32.3	36.2	5.4	-4.2	14.9
Child race/ethnicity						
White (ref)	111	27.5	32.8			
Hispanic	3	16.7	28.9	-10.8	-48.4	26.8
Black	60	25.8	28.4	-1.6	-11.9	8.7
Other**	27	44.4	40.0	17.0*	3.2	30.8
Insurance type						
Private (ref)	87	32.8	36.4			
Public/uninsured	117	28.2	31.7	-4.6	-14.0	4.9
PMCA category†						
Non-chronic (ref)	44	19.3	26.9			
Non-complex chronic	53	29.2	26.7	9.9	-0.4	20.3
Complex chronic	73	17.8	24.1	-1.5	-11.2	8.2

*p<0.05. Differences tested using linear regression.
**"Other" includes Asian/Pacific Islander, Native American, Other, and Multiracial
†PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015). Includes data from 2 children's hospitals only

~~Table 2b5.2b. Socio-Demographic Group Scores for Pediatric Psychosis: Screening for Drugs of Abuse in the Emergency Department~~						
	~~N~~	~~Mean~~	~~SD~~	~~Difference~~	~~LCL~~	~~UCL~~
~~Child gender~~						
~~Female (ref)~~	~~98~~	~~27.0~~	~~32.2~~			
~~Male~~	~~150~~	~~29.3~~	~~36.3~~	~~2.3~~	~~-6.6~~	~~11.2~~
~~Child race/ethnicity~~						
~~White (ref)~~	~~134~~	~~28.0~~	~~34.4~~			
~~Hispanic~~	~~3~~	~~16.7~~	~~28.9~~	~~-11.3~~	~~-50.0~~	~~27.4~~
~~Black~~	~~76~~	~~21.7~~	~~28.7~~	~~-6.3~~	~~-15.8~~	~~3.2~~
~~Other~~	~~32~~	~~40.6~~	~~41.0~~	~~12.6~~	~~-0.4~~	~~25.7~~
~~Insurance type~~						
~~Private (ref)~~	~~106~~	~~30.7~~	~~36.9~~			
~~Public/uninsured~~	~~142~~	~~26.8~~	~~33.0~~	~~-3.9~~	~~-12.7~~	~~4.9~~
~~PMCA category**~~						
~~Non-chronic (ref)~~	~~55~~	~~15.5~~	~~25.2~~			
~~Non-complex chronic~~	~~67~~	~~24.6~~	~~26.6~~	~~9.2*~~	~~0.1~~	~~18.2~~
~~Complex chronic~~	~~80~~	~~16.9~~	~~23.8~~	~~1.4~~	~~-7.3~~	~~10.1~~

*p<0.05. Differences tested using linear regression.
**PMCA: Pediatric Medical Complexity Algorithm (Simon et al. 2015). Includes data from 2 children's hospitals only

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?*)

For this pilot test assessing for existing variation in this measure across more than one site, we found that we were able to detect statistically and clinically meaningful differences in hospital performance. Additional information from implementation of the measure at a larger scale, as described in Section 4.1, will assist in assessing variation across a larger group of hospitals.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

Note: *This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

2b6.1. 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*)

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 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*)

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*)

Missing data likely does not contribute to substantially or meaningfully biased estimates of performance for this quality measure.

There are two potential areas for missing data: at the level of the administrative claims, which are used for sampling patients, and in the medical abstraction stage.

Administrative Claims

There are two data fields used to identify eligible patients, the diagnosis fields and the patient age. Patient age is generally considered a reliable field and has minimal missing data.

A primary diagnosis is required for billing, and therefore also is rarely missing. It is known that some providers under-code for mental health diagnoses, which would lead to a risk of under recognition of eligible cases. This may lead to difficulty in capturing reliable estimates of performance at each hospital site, but is less likely to lead to biased estimates.

Medical abstraction

Missing data in the medical abstraction stage is interpreted as the patient not meeting the metric. It would be very unusual for a laboratory test (urine or serum) to be sent, processed, and not documented, due to regulation around laboratory reporting and quality assurance, as well as the financial imperative to bill and be reimbursed for the testing. Hence, we believe it is reasonable to assume that if these data elements are missing from the health record, then the process of care was not performed. Such cases are scored as not having passed the quality measure. It is unlikely that there is a substantial incidence of false negatives due to missing data, or of biased estimates due to differentially missing data.

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*)

It was not possible to determine how often the data described above were missing. For administrative data, if a child had a diagnosis of psychosis, but this was not coded for the encounter, there would be no way to know this other than to abstract all charts for children in the eligible age range who had ED visits during the measurement timeframe to assess the frequency with which this diagnosis is documented in the record but not coded for in billing data. This approach would not be logistically feasible. For laboratory data in medical records, we believe the true rate of missing data for tests that were actually performed would be exceedingly rare for the reasons we have outlined under section 2b7.1. There would be no way to assess whether a missing lab value, where there is no evidence in the medical record of either a lab order or test result, was secondary to not doing the test versus the order and/or test result not being recorded.

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

It is unlikely that missing data contributes to substantial or meaningful biases of performance estimates.

REFERENCES

1. Brook RH. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel RA, eds. *Clinical practice guidelines development:methodology perspectives*. Rockville, MD: Agency for Health Care Policy and Research; 1994.

2. Simon TD, Cawthon ML, Stanford S, et al. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. *Pediatrics.* 2014;133(6):e1647-1654.
3. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy.* 2005;85(3):257-268.
4. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-174.
5. Lyratzopoulos G, Elliott MN, Barbiere JM, et al. How can health care organizations be reliably compared?: Lessons from a national survey of patient experience. *Med Care.* 2011;49(8):724-733.
6. Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. *N Engl J Med.* 1998;338(26):1888-1896.
7. Shekelle PG, Chassin MR, Park RE. Assessing the predictive ability of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy. *Int J Technol Assess Health Care.* 1998;14(4):707-727.
8. Kravitz RL, Park RE, Kahan JP. Measuring the clinical consistency of panelists' appropriateness ratings: the case of coronary artery bypass surgery. *Health Policy.* 1997;42(2):135-143.
9. Hemingway H, Crook AM, Feder G, et al. Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization. *N Engl J Med.* 2001;344 (9):645-654.
10. Selby JV, Fireman BH, Lundstrom RJ, et al. Variation among hospitals in coronaryangiography practices and outcomes after myocardial infarction in a large health maintenance organization. *N Engl J Med.* 1996;335(25):1888-1896.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

In development of measure specifications using sample records from the field test hospitals, we found that it was important to specify the types of laboratory tests that might be sent to test for alcohol and drugs. We document this in the data collection tool for review during abstraction, using the following language:

"Indicate if the patient had a urine or serum toxicology screen for alcohol and drugs. The alcohol test will be a separate test from the drug tests. The drug test must be comprehensive in that it tests for multiple types of illicit drugs. Do NOT give credit for tests that include results of just a single drug. Drug screens commonly include tests for benzodiazepines, barbiturates, methamphetamine, cocaine, methadone, opiates, tetrahydrocannabinol, etc."

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

No proprietary elements are used in implementing this measure. There are no licenses or fees or other requirements needed to use any aspect of the measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This measure is part of a set of mental health quality measures the COE4CCN developed as part of the Pediatric Quality Measurement Program, funded by AHRQ, using CHIPRA monies. It has not yet been implemented as the development, validation, and testing were just recently completed. The tools needed to abstract the measures are publicly available and non-proprietary, so interested parties can implement them at any time.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Children's Hospital Association (CHA) has had representation on the National Advisory Board for COECCN since its inception. CHA has shown great interest in promoting the adoption of inpatient and ED-based measures developed by our Center. The intended audience would be hospital administrators at CHA member hospitals. We would intend to work with CHA to implement these measures over the next several years.

We also intend to publish the development and field testing of these measures in peer reviewed pediatric journals over the next 12

months. Within these publications we will include the URL where the measure data abstraction tool, measure specifications, and abstractor training materials are housed promoting further access to and dissemination of the measures.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (trends in performance results, number and percentage of people receiving high-quality healthcare)**
- **Geographic area and number and percentage of accountable entities and patients included**

N/A

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
Credible rationale

The overall goal behind capturing performance results for this measure is to optimize appropriate diagnosis in a high-risk population. The danger of misdiagnosis is two-fold. On the one hand, patients with mental illness have a high incidence of co-morbid substance abuse disorders; on the other hand intoxication with drugs of abuse or alcohol, or a mixture, may present as psychotic symptoms. Treatment of psychosis without additionally treating co-morbid substance abuse can contribute to delayed and forgone treatment for a serious mental illness. Preventing this delayed or forgone treatment has the potential to improve care and long-term outcomes for a vulnerable population, given the evidence that earlier treatment can ameliorate the severity of illness for early onset schizophrenia (see Evidence form).
As experience has borne out, quality measurement efforts can drive improvements in care, whether through increasing focus on an area of care in internal audit and feedback efforts, or through reputational or financial incentive programs (e.g., CMS' public reporting or value-based purchasing programs). We anticipate that the performance results for this measure would drive improvement through similar mechanisms.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** P2_Screen_for_Tox_Appendix_FOR_SUBMISSION-635803523158179295.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Seattle Children´s Research Institute
Co.2 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-
Co.3 **Measure Developer if different from Measure Steward:** Seattle Children´s Research Institute
Co.4 **Point of Contact:** Rita, Mangione-Smith, Rita.Mangione-Smith@seattlechildrens.org, 206-884-8242-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The COE4CCN convened two expert groups to assist in the development of the Pediatric Mental Health Measures in the Hospital Setting--the Mental Health Working Group within the COE4CCN and an external panel of experts for the Delphi panel. Please see descriptions of the groups' roles in development as well as member names listed below.
I. Mental Health Working Group: This was a group of pediatric mental health and general pediatrics experts, as well as state Medicaid leadership. Reviewed secondary database analyses of prevalence of common and costly mental health diagnoses. Developed ICD9 code definitions to identify diagnoses of interest. Reviewed and edited the literature reviews conducted by COE4CCN staff. Provided content expertise during development of the detailed measure specifications and data abstraction tool. Participated in the planning and implementation of the field test as well as interpretation of the field test results.

Members of the MHWG:

Naomi S. Bardach, MD, MAS
Assistant Professor of Pediatrics and Health Policy
Department of Pediatrics
Philip R. Lee Institute of Health Policy
University of California San Francisco

Tumaini Ruker Coker, MD, MBA
Assistant Professor of Pediatrics
David Geffen School of Medicine
University of California, Los Angeles
Associate Natural Scientist
RAND, Santa Monica

Glenace Edwall, PsyD, PhD, MPP
Director, Children's Mental Health Division
Minnesota State Health Access Data Assistance Center
Minnesota Department of Human Services

Penny Knapp, MD
Professor Emeritus
Departments of Psychiatry & Pediatrics
University of California Davis

Rita Mangione-Smith, MD, MPH
Professor and Chief | Division of General Pediatrics and Hospital Medicine
University of Washington Department of Pediatrics
Director | Quality of Care Research Fellowship
UW Department of Pediatrics and Seattle Children's Hospital
Investigator | Center for Child Health, Behavior, and Development
Seattle Children´s Research Institute

Michael Murphy, EdD
Associate Professor
Department of Psychology
Harvard Medical School
Staff Psychologist
Department of Child Psychiatry
Massachusetts General Hospital

Laura Marie Prager, MD
Associate Professor of Psychiatry
Department of Child Psychiatry
Massachusetts General Hospital

Laura Richardson, MD, MPH
Professor
Department of Pediatrics and Psychiatry
Division of Adolescent Medicine
University of Washington
Investigator
Center for Child Health, Behavior, and Development
Seattle Children's Research Institute

Bonnie Zima, MD, MPH
Professor-in-Residence
Department of Psychiatry
University of California, Los Angeles
Associate Director
UCLA Health Services Research Center

Delphi panel: Reviewed the literature review and secondary database analyses as prepared by the MHWG and COE staff. Reviewed suggested indicators for face validity and content validity based on the above materials and based on member expertise in the field.

Members of the Delphi panel:

Gary Blau, PhD
Chief, Child, Adolescent and Family Branch,
Center for Mental Health Services (CMHS),
Substance Abuse and Mental Health Services Administration (SAMHSA), Rockville, MD.
Clinical Faculty, Yale Child Study Center, Yale University

Regina Bussing, MD, MSHS
Professor, Division of Child and Adolescent Psychiatry,
Department of Psychiatry, Department of Pediatrics,
and Department of Clinical and Health Psychology,
University of Florida, Gainesville, FL
Director, Florida Outreach Project for Children and Young Adults Who Are Deaf-Blind

Thomas Chun, MD, MPH
Associate Professor, Departments of Emergency Medicine and Pediatrics
Assistant Dean of Admissions
Chair, Admissions Committee
The Alpert Medical School, Brown University
Medical Staff, Department of Pediatric Emergency Medicine
Hasbro Children's Hospital

Sean Ervin, MD, PhD
Assistant Professor in Pediatrics & General Internal Medicine
Hospitalist Medicine
Head of Section- Pediatric Hospital Medicine
Wake Forest University, School of Medicine
Winston-Salem, NC

Doris Lotz, MD, MPH
Medicaid Medical Director
New Hampshire Department of Health and Human Services
Office of Medicaid Business and Policy
Instructor, Geisel School of Medicine at Dartmouth, Department of Psychiatry

Lynn Pedraza, PhD
Executive Director of Family Voices,
Albuquerque, NM

Karen Pierce, MD, DLFAPA, DLFAACAP
Clinical Associate Professor, The Feinberg School of Medicine,
Northwestern University Medical School,
Department of Psychiatry and Behavioral Sciences, Chicago, IL,
President, Illinois Academy of Child Psychiatry

Robert Sege, MD, PhD, FAAP Professor of Pediatrics, Boston University School of Medicine Director, Division of Family and Child Advocacy, Boston Medical Center Core Faculty, Harvard Injury Control Research Center Core Faculty, Harvard Youth Violence Prevention Center Gail Slap, MD, MSc Professor of Pediatrics, Department of Pediatrics, Professor of Medicine, Department of Medicine, University of Pennsylvania School of Medicine
Measure Developer/Steward Updates and Ongoing Maintenance Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Ad.7 **Disclaimers:**
Ad.8 **Additional Information/Comments:**

PSYCHOSIS

Note: There are a number of ICD9 codes that have mapped to the same ICD10 code, and one ICD9 code that mapped to 2 ICD10 codes

ICD9 used in Field test	ICD9 label	ICD10 conversion from CMS GEMS tool	ICD10 label
291.3	alcoh psy dis w hallucin	F10.951	Alcohol use, unspecified with alcohol-induced psychotic disorder with hallucinations
291.5	alcoh psych dis w delus	F10.950	Alcohol use, unspecified with alcohol-induced psychotic disorder with delusions
292.11	drug psych disor w delus	F19.950	Other psychoactive substance use, unspecified with psychoactive substance-induced psychotic disorder with delusions
292.12	drug psy dis w hallucin	F19.951	Other psychoactive substance use, unspecified with psychoactive substance-induced psychotic disorder with hallucinations
293.81	psy dis w delus oth dis	F06.2	Psychotic disorder with delusions due to known physiological condition
293.82	psy dis w halluc oth dis	F06.0	Psychotic disorder with hallucinations due to known physiological condition
295.3	paranoid schizo-unspec	F20.0	Paranoid schizophrenia
295.31	paranoid schizo-subchr	F20.0	Paranoid schizophrenia
295.32	paranoid schizo-chronic	F20.0	Paranoid schizophrenia
295.33	paran schizo-subchr/exac	F20.0	Paranoid schizophrenia
295.34	paran schizo-chr/exacerb	F20.0	Paranoid schizophrenia
295.4	schizophreniform dis nos	F20.81	Schizophreniform disorder
295.41	schizophrenic dis-subchr	F20.81	Schizophreniform disorder
295.42	schizophren dis-chronic	F20.81	Schizophreniform disorder
295.43	schizo dis-subchr/exacer	F20.81	Schizophreniform disorder
295.44	schizophr dis-chr/exacer	F20.81	Schizophreniform disorder
295.7	schizoaffective dis nos	F25.9	Schizoaffective disorder, unspecified
295.71	schizoaffectv dis-subchr	F25.9	Schizoaffective disorder, unspecified
295.72	schizoaffective dis-chr	F25.9	Schizoaffective disorder, unspecified
295.73	schizoaff dis-subch/exac	F25.9	Schizoaffective disorder, unspecified
295.74	schizoafftv dis-chr/exac	F25.9	Schizoaffective disorder, unspecified
295.9	schizophrenia nos-unspec	F20.9	Schizophrenia, unspecified
295.91	schizophrenia nos-subchr	F20.9	Schizophrenia, unspecified
295.92	schizophrenia nos-chr	F20.9	Schizophrenia, unspecified
295.93	schizo nos-subchr/exacer	F20.9	Schizophrenia, unspecified
295.94	schizo nos-chr/exacerb	F20.9	Schizophrenia, unspecified
296.24	depr psychos-sev w psych	F32.3	Major depressive disorder, single episode, severe with psychotic features
296.44	bipol i manic-sev w psy	F31.2	Bipolar disorder, current episode manic severe with psychotic features
297.1	delusional disorder	F22	Delusional disorders

297.2	paraphrenia	F22	Delusional disorders
297.3	shared psychotic disord	F22	Delusional disorders
298.0	react depress psychosis	F32.3	Major depressive disorder, single episode, severe with psychotic features (Note: This is a duplicate, with two ICD10 codes for one ICD9)
298.0	react depress psychosis	F33.3	Major depressive disorder, recurrent, severe with psychotic symptoms (Note: This is a duplicate, with two ICD10 codes for one ICD9)
298.1	excitativ type psychosis	F28	Other psychotic disorder not due to a substance or known physiological condition
298.3	acute paranoid reaction	F23	Brief psychotic disorder
298.4	psychogen paranoid psych	F23	Brief psychotic disorder
298.8	react psychosis nec/nos	F23	Brief psychotic disorder
298.9	psychosis nos	F29	Unspecified psychosis not due to a substance or known physiological condition