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: Click link to go to the link; ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3312

Measure Title: Continuity of Care for Medicaid Beneficiaries after Detoxification (Detox) From Alcohol and/or Drugs

Measure Steward: Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services

Brief Description of Measure: Percentage of discharges from a detoxification episode for adult Medicaid Beneficiaries, age 18-64, that was followed by a treatment service for substance use disorder (including the prescription or receipt of a medication to treat a substance use disorder (pharmacotherapy) within 7 or 14 days after discharge. This measure is reported across all detoxification settings.

Developer Rationale: Nearly 12 percent of Medicaid beneficiaries over age 18 have a substance use disorder (SUD) (SAMHSA, 2013) and 14 percent of newly eligible low-income adults have an SUD (Mark et al., 2015). Detoxification is a medical intervention that manages an individual safely through acute withdrawal from alcohol and/or drugs. It is widely agreed that detoxification focuses on managing acute intoxication and withdrawal from a substance but by itself is not treatment and does little to address long standing social and behavioral problems associated with substance use (McCorry et al., 2000; McLellan et al., 2005). The occurrence of detox from substances is high. Of annual admissions to substance use disorder treatment, 22% are for detoxification (detox) in inpatient hospital, residential, or outpatient settings (SAMHSA, 2015).

Many detox patients are repeat users of the service and have multiple detox episodes (Amodeo et al., 2008; Carrier et al., 2011; McCarty et al., 2000; McLellan et al., 2005). This is particularly true for those who are Medicaid eligible as they have been found to be more likely to have multiple detoxifications compared to those who are not on Medicaid (Carrier et al., 2011; Mark et al., 2006).

While detox is valued, follow-up care is critical after leaving detoxification and studies have shown that continuity of care is associated with better outcomes, although the time frame for continuity to occur differs across studies. However, research has shown that large numbers of people each year receiving detoxification services do not receive follow-up treatment (Carrier et al., 2011; Center for Substance Abuse Treatment, 2006; Specka et al., 2011).

Several studies indicate that more than half of detox patients do not receive continuity of care after detox within timeframes ranging from 14 days to six months across different studies. A study of patients in public sector substance abuse treatment in five states found continuity care within 14 days of leaving detox to range from 12.5% to 45.5%, depending on the state (Lee et al., 2014). Another study in New York State public sector treatment found a 48% continuity of care rate within six months of detox admission (Carrier et al., 2011). Data from the Integrated Database (IDB) combining administrative data from state Medicaid programs, mental health agencies, and substance abuse agencies at the client level from three states: Delaware, Oklahoma, and Washington show an overall continuity of care rate within 30 days of 27% (Mark et al., 2006).

The risks of not having continuity of care include multiple readmissions, continued criminal justice involvement, and lower employment status (Ford and Zarate, 2010; Lee et al., 2014; Mark et al., 2006; McCusker et al., 1995).

Numerator Statement: Discharges in the denominator who have an inpatient, intensive outpatient, partial hospitalization, outpatient visit, residential, or drug prescription or procedure within 7 or 14 days after discharge from a detoxification episode.

Denominator Statement: Adult Medicaid beneficiary discharges from detoxification from January 1 to December 15 of the measurement year.

Denominator Exclusions: Not applicable. The measure does not have denominator exclusions.

Measure Type: Process

Data Source: Claims

Level of Analysis: Population : Regional and State

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

Staff Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* 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. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

- In their logic model, the developer links this measure to a reduction in substance use, reduced readmissions to detox, longer time to readmission, reduction in criminal justice involvement, improved employment status, and a reduction in healthcare costs.
- In addition to the rationale, the developer also cited 15 studies and articles to support the measure:
    - Individuals who experience detox not followed by rehabilitative treatment are likely to relapse.
    - Patients who entered treatment within 3 days of inpatient detox discharge were less likely to have repeat crisis detox visits than those who did not enter treatment.
    - Patients who entered treatment within 14 days of detox discharge were less likely to be readmitted to detox
- This measure is proposed as a result of an extensive environmental scan through PubMed searches as well as through a TEP that was convened.

Questions for the Committee:

- *Does the empirical evidence submitted summarize all studies in the body of evidence?*
- *What is the relationship of this measure to patient outcomes?*

- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*
- *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*

Guidance from the Evidence Algorithm

Empirical evidence submitted without systematic review (Box 7) → Includes all studies in the bodies of evidence (Box 8) → evidence indicates high certainty that benefits outweigh undesirable effects (box 9) → Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

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 provided 2013-2014 Medicaid data to look at overall continuity performance rates for 7-day and 14-day follow-up for all states:
  - 14-day results
    - Population: 67,719
    - Performance rate: 36.5%
  - 7-day results
    - Population: 67,719
    - Performance rate: 28.8%

Disparities

- The 7-day and 14-day measure performance was stratified for disparities by age, race, and ethnicity. Testing results across all states showed higher continuity rates for white beneficiaries, females, and those aged 18-24.
- 14-day continuity by age (all states)

Age	Performance Rate
18-24	43.2%
25-44	38.5%
45-64	32.2%

- 14-day continuity by gender (all states)

Gender	Performance Rate
Male	33.9%
Female	41.2%

- 14-day continuity by race/ethnicity (all states)

Race/Ethnicity	Performance Rate
White	41.3%
Black	32.3%

Race/Ethnicity	Performance Rate
American Indian/Alaskan Native	32.3%
Asian	28.9%
Hispanic/Latino	28.5%
Native Hawaiian/Pacific Islander	27.3%
Other	31.3%
Unknown	27.8%

- 7-day continuity rate by age (all states)

Age	Performance Rate
18-24	37.2%
25-44	30.8%
45-64	24.0%

- 7-day continuity rate by gender (all states)

Gender	Performance Rate
Male	25.9%
Female	34.1%

- 7-day continuity rate by race/ethnicity (all states)

Race/Ethnicity	Performance Rate
White	33.6%
Black	24.3%
American Indian/Alaskan Native	26.4%
Asian	19.1%
Hispanic/Latino	20.7%
Native Hawaiian/Pacific Islander	20.5%
Other	25.4%
Unknown	21.4%

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence

Comments:

** There are studies linking earlier follow up (I am loath to call this continuity) to enhanced outcomes These data are largely historical observational data and thus the strength of evidence is limited.

** The data used to support this measure is from a literature search. The evidence suggests that more proximal follow-up with treatment following detox results in reduced morbidity (repeat detox, reduced co-morbidity) and improved quality of life (increased employment rate, decreased jail time). It is not clear to me that the studies cited use the same definition of follow-up (e.g. include medication in addition to other treatment modalities)

** Although the evidence to support improved outcomes for individuals receiving follow up care post detox is very strong, the evidence cited is a bit dated, 2014 is the latest study. This is at the start of the rapid rise in the opiate epidemic and also predates a shift in practice to treat SUD in primary care settings.

** The measure uses early follow up after detox as a proxy for continuity of care. they do not cite a systematic review or guidelines but do cite several individual papers that suggest that earlier follow up or better continuity of care relates to better health and social outcomes.

** evidence to support measure focus is strong

** Agree this is process measure targeting to decrease readmission/increase sobriety rates post detox, evidence is not a systematic review, yet relevant to argument and is accurate--> moderate rating is appropriate

**evidence exists; review of articles. but i'm confused about why '7 day or 14 day,' and if this is the best way of determining improved outcomes. i imagine this population often gets lost to follow up, and if more focus should be placed on reaching out to encourage follow-up and ongoing treatment participation.

** Developers provide evidence through a literature review of 15 studies that show that persons that receive treatment within 14 days of detox experience better outcomes in terms of reduced substance use, readmissions to detox, and criminal justice involvement. Further evidence also suggests that improved continuity of care leads to improved employment post treatment and overall lower healthcare costs. There does appear to be sufficient evidence suggesting that the relationship between time from existing detox to entering SUD treatment can result in improved patient-level and system-level outcomes. Direct relationships between the process of care and outcomes exist between reduction in substance use, readmissions to detox, and healthcare costs. Indirect relationships seem to be more likely for criminal justice involvement and employment status.

** Developers provide evidence through a literature review of 15 studies that show that persons that receive treatment within 14 days of detox experience better outcomes in terms of reduced substance use, readmissions to detox, and criminal justice involvement. Further evidence also suggests that improved continuity of care leads to improved employment post treatment and overall lower healthcare costs. There does appear to be sufficient evidence suggesting that the relationship between time from existing detox to entering SUD treatment can result in improved patient-level and system-level outcomes. Direct relationships between the process of care and outcomes exist between reduction in substance use, readmissions to detox, and healthcare costs. Indirect relationships seem to be more likely for criminal justice involvement and employment status.

**I think this measure has merit – it is voluntary but will allow states to think about how to gather this information and overcome the obstacles that will arise.

**There is enough evidence to move ahead at this time with this measure.

**There is strong evidence for the measure

1b. Performance Gap

Comments:

** Yes. I would have liked to see more analysis of disparities and populations at increased risk (e.g., homeless). For state wide analyses this may not be as important unless different benefits/enrollment criteria are present.

** The measure developer used data from 14 states and demonstrated variability in performance. Even the best performing states had room for improvement.

** There is a significant gap in care. Many individuals that receive detox services do not receive follow-up SUD treatment and relapse.
** gaps and disparities were shown.
** gap in performance demonstrated.
** Agree, clearly a gab and significant area of need nationally, would have been curious to see variation by state.
** gap demonstrated in medicaid population
** Large number of individuals receiving detoxification services do not receive follow-up treatment. Studies suggest that over 50% of patients do not receive continuity of care after detox within times ranging from 14 days to 6 months. There appears to be sufficient performance variation between states on this measure. There also appears to be performance variation on this measure with regards to race/ethnicity, age and gender.
** The Gap is significant.
** May want to rename the measure as attendance at one visit post-discharge does not equal true continuity of care. Measure should include information and assistance to encourage health care systems to do more to ensure that patient attends the appointment as well as goes to visits past this one.
** Current performance was provided. Evidence showed the gap in care for follow up and also increased utilization when there was not follow up.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability**;** Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Evaluation of Reliability and Validity: Link A

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions)?*

Preliminary rating for reliability:	☒ **High**	☐ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☒ **High**	☐ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability

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.**

Measure Number: 3312

Measure Title: Continuity of Care for Medicaid Beneficiaries after Detoxification (Detox) from Alcohol and/or Drugs

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?

   NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

   TIPS: Consider the following: 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?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

   The dataset included Medicaid Analytic eXtract (MAX) 2013 and 2014 data for eligible, inpatient, other services, long-term care, and drug files. A total of 14 states and 47,313 beneficiaries were included in the testing.

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒Yes (go to Question #6)

   ☐No (please explain below then go to Question #8)

   The developer used a beta-binominal model to assess the signal-to-noise ratio. Results of reliability testing was 0.98 across states.

Continuity Threshold	Average reliability score	Range of reliability scores
7-day	0.99	(0.99-0.99)
14-day	0.99	(0.98-0.99)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

   Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☒High (go to Question #8)

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☒Yes (go to Question #9)

   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☒Yes (go to Question #10)

   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

   The developer used Spearman's correlation across four quarters of the 2014 measurement year. The results show high (at or above 0.90) temporal stability of the measure over time.

Continuity Threshold	Average across 4 quarters	Q1 vs. Q2	Q2 vs. Q3	Q3 vs. Q4
7-day	0.93	0.94	0.94	0.92
14-day	0.93	0.96	0.92	0.92

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒Yes (go to Question #2)

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

   ☐Yes (please explain below then go to Question #3)

   ☐No (go to Question #3)

   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

This is a process measure.

a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

b. Are social risk factors included in risk model? ☐Yes ☐No

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☐Yes (please explain below then go to Question #4)

☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

☐Yes (please explain below then go to Question #5)

☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

☐Yes (please explain below then go to Question #6)

☐No (go to Question #6)

☒Not applicable (go to Question #6)
Measure not specified for more than one data source.

6. Analysis of potential threats to validity: Any concerns regarding missing data?

☒Yes (please explain below then go to Question #7)

☐No (go to Question #7)

The developer noted a few issues with missing data:

- Measure utilizes secondary procedure and diagnosis codes but in the validation and anomaly tables, missing information is documented only for primary diagnosis code and "a" procedure code. This absence of secondary primary and procedure codes may reflect the beneficiaries true clinical diagnosis.
- Lack of a primary diagnosis code in the OT file and procedure code in the IP and OT files varied by state.
- New York and New Jersey use state-specific codes for methadone treatment claims
- New York frequency uses state-specific procedure claims.

The developer advises measure implementers to include the relevant state-specific codes in the measure specification calculation.

ASSESSMENT OF MEASURE TESTING

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

   ☐No (please explain below then go to Question #8)

   To assess the validity of the measure, the developer examined the association between presence and absence of continuity of care. The convergent validity of the measure had a lower odd of readmission to detox or overdose treatment among detox episodes with continuity (8.3% lower for those with continuity vs. those without).

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

   ☐Yes (go to Question #9)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☒Yes (go to Question #11)

    ☐No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☒Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)
    The developer compared performance across state-level continuity rates to understand the variation in performance. They examined the distribution of the measure across states, calculated the 95% confidence interval of the continuity rates for each state using a z-distribution for proportion. Finally, they compared each state's confidence interval to the overall measure rate that uses all beneficiaries across states.

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☒High (go to Question #14)

    ☐Moderate (go to Question #14)

    ☐Low (please explain below then go to Question #13)

    ☐Insufficient

    The 14-day continuity measure rate indicated that 6 states have a measure rate significantly greater than the overall measure rate, 5 states have a measure rate significantly lower than the overall measure rate, and the remaining 3 states had measure rates which were indistinguishable from the overall measure rate.

    The 7-day continuity measure rate indicated that 7 states have a measure rate significantly greater than the overall measure rate, 5 states have a measure rate significantly lower than the overall measure rate, and the remaining 2 states had measure rates which were indistinguishable from the overall measure rate.

13. Was other validity testing reported?

    ☐Yes (go to Question #14)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

    TIPS: Prior validity studies of the same data elements may be submitted

    ☒Yes (go to Question #15)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☒Yes (go to Question #16)

    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

Comments:

** ok by me

** I believe that gabapentin is now being used as a pharmacotherapy and is not included in the specifications. I also was unclear why inpatient admission after day of discharge was also included in the numerator (?this is what is trying to be reduced).

** The numerator does not seem to include primary care as a SUD treatment setting--although the value sets seem to include other than licensed SUD treatment providers.

The term "drug prescription of procedure" is not well defined. Is this referring to pharmacotherapy? if so, it requires greater specificity.

Also, the denominator does not seem to include overdose treatment like Narcan reversals conducted in the ED.

I am not clear why 14 day follow-up care was selected rather than 7 day follow up care.

** reliability seems well shown.

** specifications are clear.

Is methadone included in pharmacotherapy options?

injectable long acting naltrexone is mentioned, but are other medications that combine medication types, like buprenorphine and ntx, included?

Lee JD, et al Comparative effectiveness of extended release naltrexone versus buprenorphine-naltrexone for opioid relapse prevention: a multicenter, open-label, randomized controlled trial. Lancet Nov 14, 2017

To double check, methadone can only be dispensed in licensed opioid treatment programs-- are these captured in the Medicaid claims data?

** Curious about codes used to confirm treatment or receipt of medications (presumably MAT)

** All codes and descriptors included. No issues with exclusions.

** These are reasonable. One might quibble about the use of samples but I don't think they would significantly alter the results.

** This measure needs to be applied regardless of co-morbid diagnoses and not have those serve as exclusions. Additional information for heard to reach populations may need to be explored (homeless, veterans)

** Codes are specific - Question though - Need to check my state i.e. NY as we submit APG codes for Medicaid and may not match the HCPCS codes. Other states might use other codes as well and will need to provide a xwalk. I have a question out for clarification now.

2a2. Reliability – Testing

Comments:

** Depends on the unit of analysis. While this measure may be reliable at a state level I have concerns at smaller units of analysis.

** They used signal to noise analysis and temporal consistency across four quarters of the measurement year for each state. High reliability was demonstrated.

** The data included Medicaid claims from 14 states including NJ from 2013 or 2014. In NJ all SUD treatment services do not appear in the medicaid claims data because they aer carved out of the Medicaid program.

** reliability seems well shown.

** based solely on signal to noise ratio however appropriate for data source.
** no
** no concerns; signal to noise data provided.
** Data can easily be extracted from Medicaid claims data across all states and can be consistently implemented. High noise-to-noise reliability reported.
** no
** Wonder if claims will be captured accurately – is there risk of people not being counted – either in the numerator or denominator and there being underreporting? I don't feel that is a barrier regardless.
** Should treatment with pharmacotherapy alone be considered sufficient for the numerator? Seems possible that if person not connected to a provider for long-term care, this isn't sufficient alone.
** Might be an issue with timing of claims submissions. Could be a situation where on 6 months of claims are submitted and processed by the end of the year.

2b1. Validity –Testing

2b4-7. Threats to Validity

2b4. Meaningful Differences

Comments:

** I believe validity is most threatened by disparities and differences in the populations measured. To take an extreme, take one population of homeless veterans and another of upper income persons referred vie an EAP. Apples and oranges... I also worry about data completeness with the variety of differing carve outs and managed care options.
** Empirical validity testing was used. Convergent validity was acceptable (8.3% lower odds for readmission to detox or overdose treatment) among detox episodes with 14 day continuity of treatment.
** validity seems well shown. relationship of earlier follow up to lower likelihood of readmission or overdose supports validity and also, despite the fact that earlier follow up is an imperfect proxy of continuity of care, supports the evidence to support measure focus.
** I would assume how problems with missing data would be addressed in the data analysis and reported when reporting measure adherence
** some questions about which medications codes are being included, is it drawing from behavioral health side of state Medicaid only vs including primary care side as well? Things like MAT would bridge both systems for states who separate BH and PC. Maybe missing some treatment data.
** No.
** I don't think these issues are show stoppers.
** Follow-up 7 or 14 days after discharge is not enough to ensure better outcomes. But this is a step in the right direction for patients getting lost and falling through cracks in the health care system.
** There should be no distinction whether the patient has substance use as a primary or secondary diagnosis as these are often arbitrary.
** I also believe that all settings should be required to incorporate this measure (ED, inpatient hospital, psychiatric hospital, primary care). Can this also include urgent care and other drop-ins?
** As above timeliness of claims data might be an issue

2b2-3. Other Threats to Validity

2b2. Exclusions

2b3. Risk Adjustment:

Comments:

** See above. Also depends heavily on the unit of analysis.
** There were some challenges in terms of variability of coding among different states that would need to be worked out prior to measure implementation. Even so, the results remain valid per data demonstrated by measure developer.
** No risk adjustment or risk stratification.
** No issues.

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 required clinical data elements are routinely generated and used during care delivery
- All data elements are in defined fields in electronic claims
- The developer did not include any difficulties regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, or time and cost of data collection.
- No fees or licensure requirements are required.

Questions for the Committee:

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

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

Comments:

** Carve-outs threaten the feasibility of garnering complete data.
** Feasible--all data elements in defined fields in electronic claims.
** For State Medicaid systems with MH and SUD services cared out detox, overdose treatment, and SUD counseling admissions do not always appear in the Medicaid claims data.
** Because this is a claims-based measure, it seems feasible to measure and report.
** highly feasible
** yes, required elements routinely generated and additional data collection strategies are not required.
** data gathering appears feasible to obtain.
** All data elements are routinely generated and included in claims data. All elements are available electronically. This would be an easy measure to put into operational use.
** It's feasible.
** Who will this measure be counted "against" – the agency that was supposed to see the individual or the health care from where the patient was released? In what way can this encourage collaborative care?

Criterion 4: Usability and Use

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☐ **Yes** ☒ **No**

Accountability program details

- This is a new measure, so it has not been used yet.
- The developer states that this measure is intended for voluntary use by states to monitor and improve the quality of care for Medicaid beneficiaries with SUD.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- N/A. This is a new measure and has not been implemented as of yet.

Additional Feedback

- N/A

***Questions for the Committee**:*

- o *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

- The developer expects adoption of this measure to improve the quality of care for Medicaid beneficiaries who have a SUD after they are discharged from detox for alcohol and/or drugs.

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation

- None reported by the developer

Potential harms

- None reported by the developer

Additional Feedback:

- This is a new measure and has not been implemented as of yet.

***Questions for the Committee**:*

- o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- o *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for Usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE:

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:
** ok
** New measure; anticipate it will be used by states to monitor and improve quality of care provided for Medicaid beneficiaries with alcohol related use disorders.
** Which entity is held accountable for the continuity of care? The health plan? The detox agency, the SUD treatment provider? What about the role of primary care?
** The measure is not currently being used.
** measure would help push systems toward investing in post detox treatment options, however, these are usually heavily dependent on state regs/financing/billing, so it's likely states would need to be able to work w/ healthcare systems to implement change before outcomes could be improved. Insufficient treatment programs or capacity is usually largest barrier and lack of state financial support usually most signifiant limiting factor in increasing the capacity for SUD treatment.
** i question if this measure represents the best/most meaningful way to necessarily associate information gathered with improved outcomes for patients. but if only used for data gathering, i think it's reasonable.
** Results can be used to improve the efficiency of treatment for SUD, which has been shown to improve outcomes. Results can also be used to identify areas of needed improvement within a system. It may point to areas of a state that may have capacity issues, and result in a policy/program response to address the need/gap in treatment.
** These two should be OK.
** There will likely be considerable variability in how this is measured state by state but as a voluntary measure, this should improve state's awareness of their errors, provide opportunities for QI and accountability
** Similar measures are in place now

4b1. Usability – Improvement

Comments:
** Seems likely to have benefits outweigh harms even if there are difficulties with reliability and validity, especially used as a PI measure. Not so sure about an accountability measure.
** The measure intent to foster improved continuity of care seems like it does have potential to favorably impact care of this population.
** The measure could be used by states to improve the care of people with SUD and could be used by health care systems or providers for benchmarking.
** need to clarify types of pharmacotherapy included and how this measure will be updated to align with new medication treatment to further assess any unintended consequences....
Will this measure capture of treatment programs that vary widely in licensure requirements and capacity to even provide medication treatment?
** none expected
**i don't see any specific harms, other than equating gathering data for follow up as defined with necessarily improved outcomes.
** As long as the measure is used to improve quality and efficiency, and not used in a punitive way, this measure could result in better outcomes for patients. Getting individuals into treatment more efficiently is always preferable.
** I wonder if the complete emphasis on meds might have the unintended consequence of devaluing or/or decreasing the utilization of motivational interviewing and harm reduction approaches for patients that are not yet willing to start formal treatment but wanting to do better.
** Possible that one visit post discharge will imply strong continuity of care even though patient may not return to the practitioner again. Could have unintended consequence of suggesting health care provider doing all possible to keep patient in care even though should be health care system's responsibility to educate and inform patient and family about long-term care, next steps, what to expect, treatment adherence, etc. May not reduce recidivism as intended. I don't think harm outweighs the good and this is a step in right direction. Though I think this measure could be bolstered by some additional language and suggestions to improve.
** So now there will be several measures out that do basically the same thing. #2605 and this one. That gets confusing to providers. It would help if NQF approved one or the other and not both.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0004: Initiation and Engagement of Alcohol and Other Drug Dependence Treatment (IET)
- 2605: Follow-Up After Emergency Department Visit for Mental Illness or Alcohol and Other Drug Dependence

Harmonization

- The developer stated that #2605 examines follow-up care 7 days and 30 days after discharge while #3312 examines follow-up care 7 days and 14 days after discharge.
- The developer stated that #2605 includes outpatient visits, intensive outpatient visits or partial hospitalizations and #3312 includes the same locations as 2605 in addition to pharmacotherapy on day of discharge, residential treatment, and long-term care. These additional follow-up services are all valid and appropriate treatments after detoxification. If the developer were to exclude them from the measure specifications, measure implementers would undercount the number of beneficiaries receiving continuity of care after detoxification.
- The developer stated that #2605 requires a primary diagnosis of alcohol and other drug dependence (AOD) for the follow-up service while #3312 requires a primary or secondary diagnosis of AOD. #3312 allow a primary or secondary AOD diagnosis to address potential inaccuracies in how AOD diagnoses are coded. For example, some providers may be concerned about the stigma associated with an AOD diagnosis and therefore code it as a secondary diagnosis. Also, for adults with co-occurring mental health and AOD disorders, the assignment of primary and secondary diagnoses can be challenging and sometimes arbitrary.
- The differences in measure specifications between 2605 and 3312 are minor and expected to have minimal impact on data collection burden.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January 10, 2018

- No comments received.
- Of the 1 NQF member who submitted a support/non-support choice:
    1. 1 supports the measure
    2. 0 do not support the measure

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

Cont_Care_After_Detox__Evidence_Attachment.pdf

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a Evidence (subcriterion 1a)

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

Measure Title: **Continuity of Care for Medicaid Beneficiaries after Detoxification (Detox) From Alcohol and /or Drugs**

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 10/31/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- 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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☐ 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. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

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

☒ Process: Continuity of Care for Medicaid Beneficiaries after Detoxification (Detox) From Alcohol and /or Drugs

☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Continuity of care after detox has been low across multiple sites, and therefore, endorsement and implementation of a measure may be used for public reporting to drive quality improvement. The potential benefits of such a measure at the client level include reduction in substance use (McCusker, Bigelow, Luippold, Zorn, & Lewis, 1995; McLellan, Weinstein, Shen, Kendig, & Levine, 2005),

readmission to detox (Carrier et al., 2011; Ford & Zarate, 2010; Lee et al., 2014; Mark, Vandivort-Warren, & Montejano, 2006), and criminal justice involvement (Ford & Zarate, 2010).

The impact on clients also includes longer time to readmission and improved employment status (Ford & Zarate, 2010). The benefits to society include a reduction in costs related to criminal activity, and a reduction in healthcare costs as a result of fewer repeat detoxes (Alexandre et al., 2012; Kertesz, Horton, Friedmann, Saitz, & Samet, 2003; McCollister & French, 2003; McCollister, French, & Fang, 2010).

Factors that influence the measure are data availability and completeness and systems limitations. The former relates to the quality of data that reporting entities have available to them for calculating the measure, while the latter relates to system capacity and whether there is an adequate number of treatment slots to transition patients to as they are discharged from detox.

The logic model below also shows costs of implementing and costs of maintaining the measure as well as potential unintended consequences.

Logic model for continuity of care for Medicaid beneficiaries after detoxification from alcohol and/or drugs

Measure Information

Measure Description: Continuity of Care for Medicaid Beneficiaries After Detoxification (Detox) From Alcohol and/or Drugs (SUD-5)

Numerator: Discharges in the denominator who had an inpatient, intensive outpatient, partial hospitalization, or outpatient visit within 14 days after discharge from an inpatient hospital, residential addiction program, or ambulatory detoxification (withdrawal management).

Denominator: Adult Medicaid beneficiary discharges from detoxification from January 1 to December 15 of the measurement year. The measure will be calculated both overall and also stratified by location of detoxification in hospital inpatient, residential addiction treatment program, or ambulatory settings.

Exclusions:

- Medicaid beneficiaries with a gap in enrollment during the 14 day follow-up period after detoxification
- Medicaid beneficiaries with a medical (not SUD-associated) acute inpatient admission during the 14 day follow-up period after detoxification

Measure Uses

Public reporting

Quality Improvement (Internal to the specific organization) Supports internal quality monitoring and improvement at the program, provider, or health plan level.

Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

Benefits

Health Care:

- Across multiple studies, continuity after detox is below 50%; therefore there is much room for improvement in connecting detox clients to treatment.

Health Outcomes:

Impact on clients

- Reduction in substance use
- Reduced readmission to detox, especially detox that is not followed with any treatment
- Longer time to readmission
- Reduction in criminal justice involvement
- Improved employment status

Impact on society

- Reduction in costs related to criminal activity

Health Care Costs:

- Reduction in cost as a result of fewer repeat detoxes that are not connected to continuing treatment

Costs

Implementation Costs:

- Low cost to adopt measure because it uses administrative data
- Time for staff to add the measure to their current set of measures
- Cost for programmers to review specifications and add coding to current programs
- Cost for additional staff to oversee transition of clients to next level of care
- Expansion of system capacity so there are enough treatment slots to transition detox in a timely manner

Intervention Costs:

- Increased cost to Medicaid if more individuals continuing onto care after detox. Cost will depend on level of care of the continuity service.
- Cost to client for additional treatment, lost days of work to enter treatment, transportation costs, and possibly child care costs.

Unintended Consequences:

- If there is not enough detox capacity, facilities could avoid clients they consider less likely to go to follow-up care.
- To meet the measure of a treatment service within 14 days, providers may pay less attention to finding an appropriate placement.
- Programs may be held accountable for events that are not under their control, e.g., client motivation or system capacity.

Influencing Factors

- Data availability and completeness
  - States' coverage of SUD treatment services
  - Medicaid systems organization
  - Coding of detoxification in claims and encounter data
- Health system issues
  - System capacity
  - Extent of network for referrals

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

**RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) **

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

Not applicable. Not an outcome measure.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Not applicable. Evidence is not based on a systematic review

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 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.

This measure is proposed as a result of an extensive environmental scan. We performed an environmental scan of existing SUD quality measures and those under development, identified the major gaps in SUD quality measurement, and recommended measure concepts and domains to fill these gaps through development and testing of de novo or adapted measures. The scan included a targeted literature review and interviews with key stakeholders representing states, managed care organizations, researchers, consumers, and providers, to identify the most promising and meaningful measures for the Medicaid program.

Throughout the process, we were guided by the priorities outlined in the SAMHSA National Behavioral Health Quality Framework (NBHQF).[1] The NBHQF goals reflect an effort to harmonize and prioritize measures that reflect the core principles of SAMHSA, as well as support the CMS National Quality Strategy.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

Not getting patients into treatment after detox is a missed opportunity to connect them to the treatment system and studies show that having continuity of care after detox is associated with the better outcomes of longer time to repeat detox and fewer detox readmissions. Individuals who experience detoxification not followed by rehabilitative treatment are likely to relapse to substance use, which may result in readmission to another detox (McLellan et al., 2005). Patients who entered treatment within three days of inpatient detox discharge were less likely to have repeat

[1] SAMHSA (2015a). National Behavioral Health Quality Framework. Retrieved December 9, 2015, from http://www.samhsa.gov/data/national-behavioral-health-quality-framework

crisis detox visits than those who did not enter treatment (Carrier et al., 2011). Another study (Lee et al., 2014) using administrative data from five states found that patients who entered treatment with 14-days of detox discharge were less likely to be readmitted to detox and that continuity of care was particularly effective in reducing readmissions to another detox that was not followed with treatment. A longer period of time between detox admissions is generally viewed as a better outcome, since it indicates that the individual is experiencing a longer period before a relapse occurs. Several studies have reported that time to readmission was longer when the client continued to treatment after detox (Mark et al., 2006; Thakur, Hoff, Druss, & Catalanotto, 1998).

Improved employment status, reduced criminal activity, and longer periods of abstinence have also been found in patients who had continuity of care after detox (Ford & Zarate, 2010). Those who entered treatment after detox were over four times as likely to being employed within three months of discharge as well as fewer arrests and fewer days in jail. Furthermore, patients who are followed up with treatment after detox have lower rates of drug use at follow-up (McCusker et al., 1995). In spite of the support that continuity of care within a short window of time after leaving detox is related to better outcomes, many do not continue onto care (Campbell et al., 2010; Carrier et al., 2011; Carroll, Triplett, & Mondimore, 2009; Mark, Dilonardo, Chalk, & Coffey, 2003; Mark et al., 2006; Stein, Kogan, & Sorbero, 2009).

1a.4.2 What process was used to identify the evidence?

PubMed searches were conducted using keywords: detox, withdrawal management, continuity of care, and outcomes. In addition, titles of key articles or author names were entered into Google Scholar to identify related articles.

In addition, a technical expert panel (TEP) was convened to discuss development of this measure on April 18, 2016, April 20, 2016, and September 21, 2016. The members of the panel advised as to which measures, from a list of measures that were found through an environmental scan as filling a measurement gap, they perceived to be important to develop and test. The TEP rated the continuity of care after detox measure high priority for development. The list of TEP members is included in Appendix A.

1a.4.3. Provide the citation(s) for the evidence.

Alexandre, P. K., Beulaygue, I. C., French, M. T., McCollister, K. E., Popovici, I., & Sayed, B. A. (2012). The economic cost of substance abuse treatment in the state of Florida. *Eval Rev, 36*(3), 167-185. doi: 10.1177/0193841X12450164

Campbell, B. K., Tillotson, C. J., Choi, D., Bryant, K., DiCenzo, J., Provost, S. E., . . . McCarty, D. (2010). Predicting outpatient treatment entry following detoxification for injection drug use: the impact of patient and program factors. *J Subst Abuse Treat, 38 Suppl 1*, S87-96. doi: S0740- 5472(10)00024-3 [pii]10.1016/j.jsat.2009.12.012

Carrier, E., McNeely, J., Lobach, I., Tay, S., Gourevitch, M. N., & Raven, M. C. (2011). Factors associated with frequent utilization of crisis substance use detoxification services. *J Addict Dis, 30*(2), 116- 122. doi: 936254277 [pii]10.1080/10550887.2011.554776

Carroll, C. P., Triplett, P. T., & Mondimore, F. M. (2009). The Intensive Treatment Unit: A brief inpatient detoxification facility demonstrating good postdetoxification treatment entry. *J Subst Abuse Treat, 37*(2), 111-119. doi: S0740-5472(08)00216-X [pii]10.1016/j.jsat.2008.11.003

Ford, L, & Zarate, P. (2010). Closing the gaps: The impact of inpatient detoxification and continuity of care on client outcomes. *Journal of Psychoactive Drugs, SARC Supplement 6, September*, 303- 314.

Kertesz, S. G., Horton, N. J., Friedmann, P. D., Saitz, R., & Samet, J. H. (2003). Slowing the revolving door: stabilization programs reduce homeless persons' substance use after detoxification. *J Subst Abuse Treat, 24*(3), 197-207.

Lee, M. T., Horgan, C. M., Garnick, D. W., Acevedo, A., Panas, L., Ritter, G. A., . . . Reynolds, M. (2014). A performance measure for continuity of care after detoxification: relationship with outcomes. *J Subst Abuse Treat, 47*(2), 130-139. doi: 10.1016/j.jsat.2014.04.002

Mark, T. L., Dilonardo, J. D., Chalk, M., & Coffey, R. (2003). Factors associated with the receipt of treatment following detoxification. *J Subst Abuse Treat, 24*(4), 299-304. doi: S0740547203000394 [pii]

Mark, T. L., Vandivort-Warren, R., & Montejano, L. B. (2006). Factors affecting detoxification readmission: analysis of public sector data from three states. *J Subst Abuse Treat, 31*(4), 439- 445. doi: S0740-5472(06)00161-9 [pii]10.1016/j.jsat.2006.05.019

McCollister, K. E., & French, M. T. (2003). The relative contribution of outcome domains in the total economic benefit of addiction interventions: a review of first findings. *Addiction, 98*(12), 1647- 1659.

McCollister, K. E., French, M. T., & Fang, H. (2010). The cost of crime to society: new crime-specific estimates for policy and program evaluation. *Drug Alcohol Depend, 108*(1-2), 98-109. doi: 10.1016/j.drugalcdep.2009.12.002

McCusker, J., Bigelow, C., Luippold, R., Zorn, M., & Lewis, B. F. (1995). Outcomes of a 21-day drug detoxification program: retention, transfer to further treatment, and HIV risk reduction. *Am J Drug Alcohol Abuse, 21*(1), 1-16.

McLellan, A. T., Weinstein, R. L., Shen, Q., Kendig, C., & Levine, M. (2005). Improving continuity of care in a public addiction treatment system with clinical case management. *Am J Addict, 14*(5), 426-440. doi: M165J02445LNQ613 [pii]10.1080/10550490500247099

Stein, B. D., Kogan, J. N., & Sorbero, M. (2009). Substance abuse detoxification and residential treatment among Medicaid-enrolled adults: rates and duration of subsequent treatment. *Drug Alcohol Depend, 104*(1-2), 100-106. doi: S0376-8716(09)00137-9 [pii]10.1016/j.drugalcdep.2009.04.008

Thakur, N. M., Hoff, R. A., Druss, B., & Catalanotto, J. (1998). Using recidivism rates as a quality indicator for substance abuse treatment programs. *Psychiatric services (Washington, D C), 49*(10), 1347- 1350.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Nearly 12 percent of Medicaid beneficiaries over age 18 have a substance use disorder (SUD) (SAMHSA, 2013) and 14 percent of newly eligible low-income adults have an SUD (Mark et al., 2015). Detoxification is a medical intervention that manages an individual safely through acute withdrawal from alcohol and/or drugs. It is widely agreed that detoxification focuses on managing acute intoxication and withdrawal from a substance but by itself is not treatment and does little to address long standing social and behavioral problems associated with substance use (McCorry et al., 2000; McLellan et al., 2005). The occurrence of detox from substances is high. Of annual admissions to substance use disorder treatment, 22% are for detoxification (detox) in inpatient hospital, residential, or outpatient settings (SAMHSA, 2015).

Many detox patients are repeat users of the service and have multiple detox episodes (Amodeo et al., 2008; Carrier et al., 2011; McCarty et al., 2000; McLellan et al., 2005). This is particularly true for those who are Medicaid eligible as they have been found to be more likely to have multiple detoxifications compared to those who are not on Medicaid (Carrier et al., 2011; Mark et al., 2006).

While detox is valued, follow-up care is critical after leaving detoxification and studies have shown that continuity of care is associated with better outcomes, although the time frame for continuity to occur differs across studies.However, research has shown that large numbers of people each year receiving detoxification services do not receive follow-up treatment (Carrier et al., 2011; Center for Substance Abuse Treatment, 2006; Specka et al., 2011).

Several studies indicate that more than half of detox patients do not receive continuity of care after detox within timeframes ranging from 14 days to six months across different studies. A study of patients in public sector substance abuse treatment in five states found continuity care within 14 days of leaving detox to range from 12.5% to 45.5%, depending on the state (Lee et al., 2014). Another study in New York State public sector treatment found a 48% continuity of care rate within six months of detox admission (Carrier et al., 2011). Data from the Integrated Database (IDB) combining administrative data from state Medicaid programs, mental health agencies, and substance abuse agencies at the client level from three states: Delaware, Oklahoma, and Washington show an overall continuity of care rate within 30 days of 27% (Mark et al., 2006).

The risks of not having continuity of care include multiple readmissions, continued criminal justice involvement, and lower employment status (Ford and Zarate, 2010; Lee et al., 2014; Mark et al., 2006; McCusker et al., 1995).

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

There is general agreement that continuity of care should occur within a short time after discharge from detox (American Society of Addiction Medicine, 2014). In response to mixed public comments regarding the follow-up time

period, we tested the measure's sensitivity to 7-day and 14-day follow-up. This approach balances clinical best practice thinking that the sooner the patient is connected to treatment the better while also allowing detox programs more time for placement of patients in follow-up treatment. Because it may be difficult at times for treatment programs to place clients in continuing care in a timely fashion after discharge due to limits in systems capacity, it is particularly important to allow more time for continuity of care to occur. The results of the additional 7-day analysis are similar to the 14-day analyses and confirm the descriptive and validity testing.

We used 2013 and 2014 Medicaid data to test the measure. Overall continuity (in inpatient, intensive outpatient, partial hospitalization, outpatient visit, residential treatment settings or through pharmacotherapy ranged from the 14-day continuity rate of 23.2 percent and 7-day continuity rate of 15.8 percent in New Jersey to the 14 day continuity rate of 84.4 percent and 7-day continuity rate of 81.4 percent in Vermont.

Below we present the total population number and performance rate, overall and for each state, for the 14-day and 7-day continuity.

14-day results

Total

Minimum: 23.15

25th percentile: 28.50

50th percentile: 38.02

75th percentile: 44.49

Maximum: 84.4

All States

-Population: 67,719

-Performance rate: 36.5%

Connecticut

-Population: 2,799

-Performance rate: 41.5%

Georgia

-Population: 1,571

-Performance rate: 26.0%

Iowa

-Population: 379

-Performance rate: 28.2%

Michigan

-Population: 3,760

-Performance rate: 67.5%

Missouri

-Population: 997

-Performance rate: 36.6%

Mississippi

-Population: 884

-Performance rate: 34.3%

New Jersey

-Population: 6068

-Performance rate: 23.2%

New York

-Population: 32,744

-Performance rate: 27.3%

Pennsylvania

-Population: 9,474

-Performance rate: 58.5%

Tennessee

-Population: 3,911

-Performance rate: 39.4%

Texas

-Population: 929

-Performance rate: 39.5%

Vermont

-Population: 1,160

-Performance rate: 84.4%

Washington

-Population: 2,058

-Performance rate: 29.3%

West Virginia

-Population: 985

-Performance rate: 45.5%

7-day results

Total

Minimum: 15.80

25th percentile: 20.97

50th percentile: 29.82

75th percentile: 37.17

Maximum: 81.38

All States

-Population: 67,719

-Performance rate: 28.8%

Connecticut

-Population: 2,799

-Performance rate: 30.7%

Georgia

-Population: 1,571

-Performance rate: 18.3%

Iowa

-Population: 379

-Performance rate: 20.6%

Michigan

-Population: 3,760

-Performance rate: 63.0%

Missouri

-Population: 997

-Performance rate: 29.0%

Mississippi

-Population: 884

-Performance rate: 27.0%

New Jersey

-Population: 6068

-Performance rate: 15.8%

New York

-Population: 32,744

-Performance rate: 18.8%

Pennsylvania

-Population: 9,474

-Performance rate: 51.5%

Tennessee

-Population: 3,911

-Performance rate: 32.9%

Texas

-Population: 929

-Performance rate: 32.7%

Vermont

-Population: 1,160

-Performance rate: 81.4%

Washington

Population: 2,058

Performance rate: 22.2%

West Virginia

Population: 985

Performance rate: 38.6%

The three states with the highest continuity rate were Michigan, Pennsylvania, and Vermont. We wanted to confirm the face validity of the rates in the three states with the highest rates; thus, we reached out to Medicaid and behavioral health agency representatives in these states. Below we have summarized the information we gathered from officials in these three states.

Michigan. A state official from the Michigan Health and Human Services confirmed that the state's continuity rate is about 66–67 percent in its urban areas.

Pennsylvania. We spoke with state officials from the Pennsylvania Department of Public Welfare, Office of Medical Assistance Programs; their reaction to our finding of on overall 58.5 percent continuity rate was that it was high. Furthermore, our findings of 9,474 detox episodes during the year seemed very low to them. They reported having many more detox episodes in a year and found that only about a third of them had continuity after detox discharge. In using the 2014 MAX data, we may have been observing only the FFS claims and gotten incomplete data for the state's managed care beneficiaries. In addition, the detox codes that we include in our testing specifications may not include all possible billing codes used in Pennsylvania.

Vermont. In a conversation with officials from Vermont, we learned that because of the way the treatment system is set up in their state, there is a high continuity rate in specialty locations, such as residential inpatient addiction treatment, where the beneficiary is admitted to detox for a couple of days and then transferred to residential treatment within the same treatment agency. Therefore, the state's continuity rates will be high in those locations. We found a relatively low number of detox episodes in hospital inpatient locations in Vermont. We learned that we were missing many of these episodes because the state does not require providers to submit detoxification ICD procedure, revenue, or HCPCS codes on billing records.

Continuity of care variations by detox location.

The overall continuity of care rate was 36.5 percent for 14-day continuity and 28.8 percent for 7-day continuity across all states tested. However, the rate varied with the location of the detox considering both 14- and 7- day continuity.

Across all states tested, for the 14-day measure, the continuity rate by detox location was 28.1 percent for hospital inpatient, 50.6 percent for ambulatory addiction treatment, 55.7 percent for other stayover, 60.8 percent for ambulatory treatment in a residential addiction specialty setting, and 62.9 percent for residential inpatient. Thus, continuity rates were lower when the detox occurred in a hospital inpatient location.

The only two exceptions were in Connecticut, where continuity of care occurred at the same rate after hospital inpatient and outpatient detox (42 percent), and in Washington, where the rate was lowest in residential inpatient (26.3 percent). In Iowa, Michigan, Missouri, New York, Pennsylvania, Tennessee, Texas, and Vermont, the lowest continuity rate occurred when detox was provided in inpatient hospitals (28.0 percent, 27.6 percent, 34.9 percent, 26.1 percent, 42.7 percent, 31.9 percent, 34.7 percent, and 40.3 percent, respectively). No state had a continuity rate higher than 46 percent when detox occurred in an inpatient hospital location.

Below we present the total population number and performance rate by location of detoxification for each state, for the 14-day continuity.

Hospital inpatient
All States
-Total: 28.1%
Connecticut
-Population: 2,799
-Performance rate: 41.5%
Georgia
-Population: 1,571
-Performance rate: 26.1%
Iowa
-Population: 379
-Performance rate: 28.0 %
Michigan
-Population: 3,760
-Performance rate: 27.6%
Missouri
-Population: 997

-Performance rate: 34.9%
Mississippi
-Population: 884
-Performance rate: 34.3%
New Jersey
-Population: 6068
-Performance rate: 23.2%
New York
-Population: 32,744
-Performance rate: 26.1%
Pennsylvania
-Population: 9,474
-Performance rate: 42.7%
Tennessee
-Population: 3,911
-Performance rate: 31.9%
Texas
-Population: 929
-Performance rate: 34.7%
Vermont
-Population: 1,160
-Performance rate: 40.3%
Washington
Population: 2,058
Performance rate: 36.1%
West Virginia
Population: 985
Performance rate: 45.5%
Residential inpatient addiction treatment
All States
-Total: 62.9%
Connecticut
-Population: 2,799
-Performance rate: 0.0
Georgia
-Population: 1,571
-Performance rate: 0.0
Iowa
-Population: 379
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Michigan
-Population: 3,760
-Performance rate: 73.9%
Missouri
-Population: 997
-Performance rate: 39.1%
Mississippi
-Population: 884
-Performance rate: 0.0
New Jersey

-Population: 6068
-Performance rate: 0.0
New York
-Population: 32,744
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Pennsylvania
-Population: 9,474
-Performance rate: 0.0
Tennessee
-Population: 3,911
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Texas
-Population: 929
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Vermont
-Population: 1,160
-Performance rate: 90.9
Washington
Population: 2,058
Performance rate: 26.3%
West Virginia
Population: 985
Performance rate: 0.0
Ambulatory treatment in specialty addiction setting
All States
-Total: 60.8%
Connecticut
-Population: 2,799
-Performance rate: 60.8%
Georgia
-Population: 1,571
-Performance rate: 0.0%
Iowa
-Population: 379
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Michigan
-Population: 3,760
-Performance rate: 58.9%
Missouri
-Population: 997
-Performance rate: 0.0%
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: Not reported; result is based on a cell size of 10 or less.

Pennsylvania
-Population: 9,474
-Performance rate: 0.0%
Tennessee
-Population: 3,911
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Texas
-Population: 929
-Performance rate: 47.8%
Vermont
-Population: 1,160
-Performance rate: 0.0%
Washington
-Population: 2,058
-Performance rate: 0.0%
West Virginia
-Population: 985
-Performance rate: 0.0%
Other stayover location
All States
-Total: 55.7%
Connecticut
-Population: 2,799
-Performance rate: 55.7%
Georgia
-Population: 1,571
-Performance rate: 0.0%
Iowa
-Population: 379
-Performance rate: 0.0%
Michigan
-Population: 3,760
-Performance rate: 0.0%
Missouri
-Population: 997
-Performance rate: 0.0%
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: 0.0%
Pennsylvania
-Population: 9,474
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Tennessee
-Population: 3,911

-Performance rate: 56.6%
Texas
-Population: 929
-Performance rate: 39.5%
Vermont
-Population: 1,160
-Performance rate: 0.0%
Washington
-Population: 2,058
-Performance rate: 0.0%
West Virginia
Population: 985
Performance rate: 0.0%
Ambulatory addiction treatment
All States
-Total: 50.6%
Connecticut
-Population: 2,799
-Performance rate: 50.6%
Georgia
-Population: 1,571
-Performance rate: 41.6%
Iowa
-Population: 379
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Michigan
-Population: 3,760
-Performance rate: 0.0%
Missouri
-Population: 997
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: 54.6%
Pennsylvania
-Population: 9,474
-Performance rate: 53.8%
Tennessee
-Population: 3,911
-Performance rate: 66.7%
Texas
-Population: 929
-Performance rate: Not reported; result is based on a cell size of 10 or less.
Vermont

-Population: 1,160
-Performance rate: 0.0%
Washington
Population: 2,058
Performance rate: 0.0%
West Virginia
Population: 985
Performance rate: 0.0%

Across all states tested, for the 7-day measure, the continuity rate by detox location was 19.7 percent for hospital inpatient, 42.6 percent for ambulatory addiction treatment, 51.2 percent for other stayover, 54.6 percent for ambulatory treatment in a residential addiction specialty setting, and 58.7 percent for residential inpatient. Thus, continuity rates for all locations were a little lower than for the 14-day measure since the 7-day measure's time frame is shorter. As with the 14-day measure, the rates were lowest when the detox occurred in a hospital inpatient location, with the exceptions of Georgia and Washington.

Below we present the total population number and performance rate by location of detoxification for each state, for the 7-day continuity.

Hospital Inpatient
All States
-Total: 19.7%
Connecticut
-Population: 2,799
-Performance rate: 29.7%
Georgia
-Population: 1,571
-Performance rate: 18.3%
Iowa
-Population: 379
-Performance rate: 20.4%
Michigan
-Population: 3,760
-Performance rate: 17.6%
Missouri
-Population: 997
-Performance rate: 25.4%
Mississippi
-Population: 884
-Performance rate: 27.0%
New Jersey
-Population: 6068
-Performance rate: 15.8%
New York
-Population: 32,744
-Performance rate: 17.6%
Pennsylvania
-Population: 9,474
-Performance rate: 31.6%
Tennessee
-Population: 3,911

-Performance rate: 24.5%
Texas
-Population: 929
-Performance rate: 27.5%
Vermont
-Population: 1,160
-Performance rate: 32.2%
Washington
Population: 2,058
Performance rate: 28.0%
West Virginia
Population: 985
Performance rate: 38.6%
Residential inpatient addiction treatment
All States
-Total: 58.7%
Connecticut
-Population: 2,799
-Performance rate: 0.0%
Georgia
-Population: 1,571
-Performance rate: 0.0%
Iowa
-Population: 379
-Performance rate: 100.0%
Michigan
-Population: 3,760
-Performance rate: 70.2%
Missouri
-Population: 997
-Performance rate: 34.5%
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: 25.0%
Pennsylvania
-Population: 9,474
-Performance rate: 0.0%
Tennessee
-Population: 3,911
-Performance rate: 80.0%
Texas

-Population: 929
-Performance rate: 100.0%
Vermont
-Population: 1,160
-Performance rate: 88.6%
Washington
Population: 2,058
Performance rate: 19.5%
West Virginia
Population: 985
Performance rate: 0.0%
Ambulatory treatment in specialty addiction setting
All States
-54.6%
Connecticut
-Population: 2,799
-Performance rate: 0.0%
Georgia
-Population: 1,571
-Performance rate: 0.0%
Iowa
-Population: 379
-Performance rate: 0.0%
Michigan
-Population: 3,760
-Performance rate: 53.3%
Missouri
-Population: 997
-Performance rate: 0.0%
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: 54.6%
Pennsylvania
-Population: 9,474
-Performance rate: 55.1%
Tennessee
-Population: 3,911
-Performance rate: 100.0%
Texas
-Population: 929
-Performance rate: 41.7%

Vermont
-Population: 1,160
-Performance rate: 0.0%
Washington
-Population: 2,058
-Performance rate: 0.0%
West Virginia
-Population: 985
-Performance rate: 0.0%
Other stayover location
All States
Total: 51.2%
Connecticut
-Population: 2,799
-Performance rate: 0.0%
Georgia
-Population: 1,571
-Performance rate: 0.0%
Iowa
-Population: 379
-Performance rate: 0.0%
Michigan
-Population: 3,760
-Performance rate: 0.0%
Missouri
-Population: 997
-Performance rate: 0.0%
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: 0.0%
Pennsylvania
-Population: 9,474
-Performance rate: 8.3%
Tennessee
-Population: 3,911
-Performance rate: 52.1%
Texas
-Population: 929
-Performance rate: 0.0%
Vermont
-Population: 1,160

-Performance rate: 0.0%
Washington
-Population: 2,058
-Performance rate: 0.0%
West Virginia
-Population: 985
-Performance rate: 0.0%
Ambulatory addiction treatment
All States
Total: 42.6%
Connecticut
-Population: 2,799
-Performance rate: 42.6%
Georgia
-Population: 1,571
-Performance rate: 34.1%
Iowa
-Population: 379
-Performance rate: 14.3
Michigan
-Population: 3,760
-Performance rate: 0.0%
Missouri
-Population: 997
-Performance rate: 0.0%
Mississippi
-Population: 884
-Performance rate: 0.0%
New Jersey
-Population: 6068
-Performance rate: 0.0%
New York
-Population: 32,744
-Performance rate: 46.5%
Pennsylvania
-Population: 9,474
-Performance rate: 40.7%
Tennessee
-Population: 3,911
-Performance rate: 57.1%
Texas
-Population: 929
-Performance rate: 63.6%
Vermont
-Population: 1,160
-Performance rate: 0.0%
Washington

-Population: 2,058
-Performance rate: 0.0%
West Virginia
-Population: 985
-Performance rate: 0.0%

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.

Not applicable. Data have been included in Section 1b.2; these data represent continuity rates from 14 states included in testing using MAX data from 2013 and 2014.

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

During testing, the 14-day and 7-day measure performance was stratified for disparities by age, race, and ethnicity. Testing results across all states showed higher continuity rates for White beneficiaries, females and those aged 18-24, although there were some differences across states.

14-day continuity
Total Performance Rate (All states): 36.5%
Performance Rate by Age (All states)
18-24
-Performance Rate: 43.2%
25-44
-Performance Rate: 38.5%
45-64
-Performance Rate: 32.2%
Performance Rate by Gender (All states)
Male
-Performance Rate: 33.9%
Female
-Performance Rate: 41.2%
Performance Rate by Race/Ethnicity (All states)
White
-Performance Rate: 41.3%
Black
-Performance Rate: 32.3%
American Indian/Alaskan Native
-Performance Rate: 32.3%
Asian
-Performance Rate: 28.9%
Hispanic/Latino
-Performance Rate: 28.5%

Native Hawaiian/Pacific Islander
-Performance Rate: 27.3%
Other Race/Ethnicity
-Performance Rate: 31.3%
Unknown Race/Ethnicity
-Performance Rate: 27.8%
Total Performance Rate (Connecticut): 41.5%
Performance Rate by Age
18-24
-Performance Rate: 39.3%
25-44
-Performance Rate: 43.4%
45-64
-Performance Rate: 40.0%
Performance Rate by Gender
Male
-Performance Rate: 40.1%
Female
-Performance Rate: 44.7%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 42.3%
Black
-Performance Rate: 35.1%
American Indian/Alaskan Native
-Performance Rate: NR
Asian
-Performance Rate: NR
Hispanic/Latino
-Performance Rate: 42.5%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 0.0%
Total Performance Rate (Georgia): 26.0%
Performance Rate by Age
18-24
-Performance Rate: 20.8%
25-44
-Performance Rate: 31.0%
45-64

-Performance Rate: 20.8%
Performance Rate by Gender
Male
-Performance Rate: 26.7%
Female
-Performance Rate: 25.5%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 29.7%
Black
-Performance Rate: 21.7%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 0.0%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 22.4%
Total Performance Rate (Iowa): 28.2%
Performance Rate by Age
18-24
-Performance Rate: NR
25-44
-Performance Rate: 36.4%
45-64
-Performance Rate: 22.5%
Performance Rate by Gender
Male
-Performance Rate: 27.5%
Female
-Performance Rate: 30.2%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 33.2%
Black
-Performance Rate: NR
American Indian/Alaskan Native
-Performance Rate: NR

Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: NR
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: NR
Total Performance Rate (Michigan): 67.5%
Performance Rate by Age
18-24
-Performance Rate: 68.6%
25-44
-Performance Rate: 68.5%
45-64
-Performance Rate: 65.7%
Performance Rate by Gender
Male
-Performance Rate: 66.2%
Female
-Performance Rate: 68.9%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 67.5%
Black
-Performance Rate: 69.3%
American Indian/Alaskan Native
-Performance Rate: 52.9%
Asian
-Performance Rate: NR
Hispanic/Latino
-Performance Rate: 61.5%
Native Hawaiian/Pacific Islander
-Performance Rate: NR
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 63.6%
Total Performance Rate (Missouri): 36.6%
Performance Rate by Age
18-24

-Performance Rate: 39.4%
25-44
-Performance Rate: 38.4%
45-64
-Performance Rate: 33.7%
Performance Rate by Gender
Male
-Performance Rate: 37.5%
Female
-Performance Rate: 35.2%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 37.8%
Black
-Performance Rate: 34.1%
American Indian/Alaskan Native
-Performance Rate: NR
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: NR
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: NR
Total Performance Rate (Mississippi): 34.3%
Performance Rate by Age
18-24
-Performance Rate: 45.9%
25-44
-Performance Rate: 35.6%
45-64
-Performance Rate: 29.0%
Performance Rate by Gender
Male
-Performance Rate: 31.5%
Female
-Performance Rate: 36.8%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 37.9%

Black
-Performance Rate: 31.2%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 0.0%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 28.2%
Total Performance Rate (New Jersey): 23.2%
Performance Rate by Age
18-24
-Performance Rate: 19.2%
25-44
-Performance Rate: 24.3%
45-64
-Performance Rate: 22.4%
Performance Rate by Gender
Male
-Performance Rate: 22.1%
Female
-Performance Rate: 24.9%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 24.4%
Black
-Performance Rate: 19.2%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 40.7%
Hispanic/Latino
-Performance Rate: 15.6%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity

-Performance Rate: 26.9%
Total Performance Rate (New York): 27.3%
Performance Rate by Age
18-24
-Performance Rate: 27.3%
25-44
-Performance Rate: 27.9%
45-64
-Performance Rate: 26.7%
Performance Rate by Gender
Male
-Performance Rate: 27.3%
Female
-Performance Rate: 27.3%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 29.7%
Black
-Performance Rate: 26.0%
American Indian/Alaskan Native
-Performance Rate: 27.9%
Asian
-Performance Rate: 28.0%
Hispanic/Latino
-Performance Rate: 26.5%
Native Hawaiian/Pacific Islander
-Performance Rate: 29.6%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 17.1%
Total Performance Rate (Pennsylvania): 58.5%
Performance Rate by Age
18-24
-Performance Rate: 60.1%
25-44
-Performance Rate: 59.6%
45-64
-Performance Rate: 54.8%
Performance Rate by Gender
Male
-Performance Rate: 57.6%
Female

-Performance Rate: 59.3%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 58.6%
Black
-Performance Rate: 60.1%
American Indian/Alaskan Native
-Performance Rate: 61.1%
Asian
-Performance Rate: NR
Hispanic/Latino
-Performance Rate: 52.8%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 0.0%
Total Performance Rate (Tennessee): 39.4%
Performance Rate by Age
18-24
-Performance Rate: 43.2%
25-44
-Performance Rate: 41.3%
45-64
-Performance Rate: 31.4%
Performance Rate by Gender
Male
-Performance Rate: 36.7%
Female
-Performance Rate: 40.9%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 41.4%
Black
-Performance Rate: 26.2%
American Indian/Alaskan Native
-Performance Rate: NR
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: NR
Native Hawaiian/Pacific Islander

-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 31.3%
Unknown Race/Ethnicity
-Performance Rate: 44.1%
Total Performance Rate (Texas): 39.5%
Performance Rate by Age
18-24
-Performance Rate: 30.2%
25-44
-Performance Rate: 44.0%
45-64
-Performance Rate: 36.6%
Performance Rate by Gender
Male
-Performance Rate: 37.9%
Female
-Performance Rate: 41.0%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 41.5%
Black
-Performance Rate: 44.2%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: NR
Hispanic/Latino
-Performance Rate: 34.3%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 39.3%
Total Performance Rate (Vermont): 84.4%
Performance Rate by Age
18-24
-Performance Rate: 87.6%
25-44
-Performance Rate: 87.8%
45-64
-Performance Rate: 65.9%

Performance Rate by Gender
Male
-Performance Rate: 82.9%
Female
-Performance Rate: 86.1%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 84.3%
Black
-Performance Rate: 93.8%
American Indian/Alaskan Native
-Performance Rate: NR
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: NR
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 81.6%
Total Performance Rate (Washington): 29.3%
Performance Rate by Age
18-24
-Performance Rate: 31.1%
25-44
-Performance Rate: 27.6%
45-64
-Performance Rate: 31.1%
Performance Rate by Gender
Male
-Performance Rate: 28.5%
Female
-Performance Rate: 29.9%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 29.3%
Black
-Performance Rate: 31.6%
American Indian/Alaskan Native
-Performance Rate: 32.0%
Asian

-Performance Rate: NR
Hispanic/Latino
-Performance Rate: 27.1%
Native Hawaiian/Pacific Islander
-Performance Rate: NR
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 27.6%
Total Performance Rate (West Virginia): 45.5%
Performance Rate by Age
18-24
-Performance Rate: 36.2%
25-44
-Performance Rate: 48.8%
45-64
-Performance Rate: 41.8%
Performance Rate by Gender
Male
-Performance Rate: 45.3%
Female
-Performance Rate: 45.8%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 46.0%
Black
-Performance Rate: NR
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 0.0%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 0.0%
Notes: Based on analysis of 2014 (2013 for Texas and Washington State) MAX PS, IP, LT, OT, and RX files.
7-day continuity
Total Performance Rate (All states): 28.8%
Performance Rate by Age (All states)

18-24
-Performance Rate: 37.2%
25-44
-Performance Rate: 30.8%
45-64
-Performance Rate: 24.0%
Performance Rate by Gender (All states)
Male
-Performance Rate: 25.9%
Female
-Performance Rate: 34.1%
Performance Rate by Race/Ethnicity (All states)
White
-Performance Rate: 33.6%
Black
-Performance Rate: 24.3%
American Indian/Alaskan Native
-Performance Rate: 26.4%
Asian
-Performance Rate: 19.1%
Hispanic/Latino
-Performance Rate: 20.7%
Native Hawaiian/Pacific Islander
-Performance Rate: 20.5%
Other Race/Ethnicity
-Performance Rate: 25.4%
Unknown Race/Ethnicity
-Performance Rate: 21.4%
Total Performance Rate (Connecticut): 30.7%
Performance Rate by Age
18-24
-Performance Rate: 26.8%
25-44
-Performance Rate: 32.5%
45-64
-Performance Rate: 29.4%
Performance Rate by Gender
Male
-Performance Rate: 29.8%
Female
-Performance Rate: 32.6%
Performance Rate by Race/Ethnicity
White

-Performance Rate: 31.3%
Black
-Performance Rate: 24.2%
American Indian/Alaskan Native
-Performance Rate: 25.0%
Asian
-Performance Rate: 30.0%
Hispanic/Latino
-Performance Rate: 32.5%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 0.0%
Total Performance Rate (Georgia): 18.3%
Performance Rate by Age
18-24
-Performance Rate: 17.0%
25-44
-Performance Rate: 22.3%
45-64
-Performance Rate: 13.6%
Performance Rate by Gender
Male
-Performance Rate: 17.6%
Female
-Performance Rate: 18.8%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 22.0%
Black
-Performance Rate: 12.8%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 0.0%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%

Unknown Race/Ethnicity
-Performance Rate: 15.3%
Total Performance Rate (Iowa): 20.6%
Performance Rate by Age
18-24
-Performance Rate: 25.0%
25-44
-Performance Rate: 27.9%
45-64
-Performance Rate: 15.0%
Performance Rate by Gender
Male
-Performance Rate: 20.5%
Female
-Performance Rate: 20.8%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 24.1%
Black
-Performance Rate: 10.3%
American Indian/Alaskan Native
-Performance Rate: 7.1%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 16.7%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 12.7%
Total Performance Rate (Michigan): 63.0%
Performance Rate by Age
18-24
-Performance Rate: 64.8%
25-44
-Performance Rate: 64.0%
45-64
-Performance Rate: 61.0%
Performance Rate by Gender
Male
-Performance Rate: 61.7%

Female
-Performance Rate: 64.4%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 62.7%
Black
-Performance Rate: 66.3%
American Indian/Alaskan Native
-Performance Rate: 41.2%
Asian
-Performance Rate: 33.3%
Hispanic/Latino
-Performance Rate: 60.0%
Native Hawaiian/Pacific Islander
-Performance Rate: 33.3%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 56.4%
Total Performance Rate (Missouri): 29.0%
Performance Rate by Age
18-24
-Performance Rate: 36.4%
25-44
-Performance Rate: 30.2%
45-64
-Performance Rate: 26.1%
Performance Rate by Gender
Male
-Performance Rate: 28.4%
Female
-Performance Rate: 29.9%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 30.1%
Black
-Performance Rate: 26.7%
American Indian/Alaskan Native
-Performance Rate: 16.7%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 40.0%

Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 15.0%
Total Performance Rate (Mississippi): 27.0%
Performance Rate by Age
18-24
-Performance Rate: 40.0%
25-44
-Performance Rate: 28.0%
45-64
-Performance Rate: 21.8%
Performance Rate by Gender
Male
-Performance Rate: 23.8%
Female
-Performance Rate: 29.9%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 31.7%
Black
-Performance Rate: 22.7%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 0.0%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 19.7%
Total Performance Rate (New Jersey): 15.8%
Performance Rate by Age
18-24
-Performance Rate: 14.7%
25-44
-Performance Rate: 16.3%
45-64

-Performance Rate: 15.3%
Performance Rate by Gender
Male
-Performance Rate: 14.8%
Female
-Performance Rate: 17.4%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 16.8%
Black
-Performance Rate: 13.1%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 18.5%
Hispanic/Latino
-Performance Rate: 10.7%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 18.3%
Total Performance Rate (New York): 18.8%
Performance Rate by Age
18-24
-Performance Rate: 20.6%
25-44
-Performance Rate: 19.6%
45-64
-Performance Rate: 17.8%
Performance Rate by Gender
Male
-Performance Rate: 18.7%
Female
-Performance Rate: 19.2%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 21.1%
Black
-Performance Rate: 16.8%
American Indian/Alaskan Native
-Performance Rate: 22.1%

Asian
-Performance Rate: 18.5%
Hispanic/Latino
-Performance Rate: 18.5%
Native Hawaiian/Pacific Islander
-Performance Rate: 22.5%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 12.2%
Total Performance Rate (Pennsylvania): 51.5%
Performance Rate by Age
18-24
-Performance Rate: 53.3%
25-44
-Performance Rate: 52.8%
45-64
-Performance Rate: 47.3%
Performance Rate by Gender
Male
-Performance Rate: 50.6%
Female
-Performance Rate: 52.3%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 51.5%
Black
-Performance Rate: 54.1%
American Indian/Alaskan Native
-Performance Rate: 55.6%
Asian
-Performance Rate: 43.8%
Hispanic/Latino
-Performance Rate: 44.9%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 52.2%
Total Performance Rate (Tennessee): 32.9%
Performance Rate by Age
18-24

-Performance Rate: 38.1%
25-44
-Performance Rate: 34.2%
45-64
-Performance Rate: 25.5%
Performance Rate by Gender
Male
-Performance Rate: 30.5%
Female
-Performance Rate: 34.3%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 34.5%
Black
-Performance Rate: 22.4%
American Indian/Alaskan Native
-Performance Rate: 50.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 33.3%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 25.4%
Unknown Race/Ethnicity
-Performance Rate: 37.3%
Total Performance Rate (Texas): 32.7%
Performance Rate by Age
18-24
-Performance Rate: 29.2%
25-44
-Performance Rate: 35.4%
45-64
-Performance Rate: 30.3%
Performance Rate by Gender
Male
-Performance Rate: 30.9%
Female
-Performance Rate: 34.4%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 33.4%

Black
-Performance Rate: 34.7%
American Indian/Alaskan Native
-Performance Rate: 0.0%
Asian
-Performance Rate: 40.0%
Hispanic/Latino
-Performance Rate: 30.2%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 33.5%
Total Performance Rate (Vermont): 81.4%
Performance Rate by Age
18-24
-Performance Rate: 86.0%
25-44
-Performance Rate: 84.3%
45-64
-Performance Rate: 62.6%
Performance Rate by Gender
Male
-Performance Rate: 79.6%
Female
-Performance Rate: 83.4%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 81.3%
Black
-Performance Rate: 93.8%
American Indian/Alaskan Native
-Performance Rate: 100.0%
Asian
-Performance Rate: 0.0%
Hispanic/Latino
-Performance Rate: 66.7%
Native Hawaiian/Pacific Islander
-Performance Rate: 0.0%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity

-Performance Rate: 78.9%
Total Performance Rate (Washington): 22.2%
Performance Rate by Age
18-24
-Performance Rate: 23.4%
25-44
-Performance Rate: 20.7%
45-64
-Performance Rate: 23.9%
Performance Rate by Gender
Male
-Performance Rate: 20.2%
Female
-Performance Rate: 23.5%
Performance Rate by Race/Ethnicity
White
-Performance Rate: 21.7%
Black
-Performance Rate: 24.8%
American Indian/Alaskan Native
-Performance Rate: 27.0%
Asian
-Performance Rate: 20.0%
Hispanic/Latino
-Performance Rate: 19.4%
Native Hawaiian/Pacific Islander
-Performance Rate: 8.3%
Other Race/Ethnicity
-Performance Rate: 0.0%
Unknown Race/Ethnicity
-Performance Rate: 23.1%
Total Performance Rate (West Virginia): 38.6%
Performance Rate by Age
18-24
-Performance Rate: 33.3%
25-44
-Performance Rate: 42.5%
45-64
-Performance Rate: 33.3%
Performance Rate by Gender
Male
-Performance Rate: 38.3%
Female

-Performance Rate: 39.1%

Performance Rate by Race/Ethnicity

White

-Performance Rate: 39.0%

Black

-Performance Rate: 19.0%

American Indian/Alaskan Native

-Performance Rate: 0.0%

Asian

-Performance Rate: 0.0%

Hispanic/Latino

-Performance Rate: 0.0%

Native Hawaiian/Pacific Islander

-Performance Rate: 0.0%

Other Race/Ethnicity

-Performance Rate: 0.0%

Unknown Race/Ethnicity

-Performance Rate: 0.0%

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

Not applicable. Performance data provided in 1b.4.

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 sub criteria 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):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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.)*

The measure does not yet have published specifications. Therefore no link exists, but specifications are attached in accordance with question S.2a.

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:** Cont_Care_After_Detox_Value_Sets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Not applicable.

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) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Discharges in the denominator who have an inpatient, intensive outpatient, partial hospitalization, outpatient visit, residential, or drug prescription or procedure within 7 or 14 days after discharge from a detoxification episode.

S.5. 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, time period for data collection, 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 (S.14).

Measure data will be reported annually (12 months). To account for the 14-day time period after discharge from detoxification, the denominator period will start January 1 and end December 15 of the measurement year.

The numerator includes individuals with any of the following within 14 days after discharge from detoxification:

-Pharmacotherapy on day of discharge through day 7 or 14.

-Outpatient, intensive outpatient, partial hospitalization, or residential

treatment procedure with a diagnosis of SUD on the day after discharge through day 7 or 14.

-Outpatient, intensive outpatient, partial hospitalization, or residential treatment with standalone SUD procedure on the day after discharge through day 7 or 14.

-Inpatient admission with an SUD diagnosis or procedure code on day after discharge through day 7 or 14.

-Long-term care institutional claims with an SUD diagnosis on day after discharge through day 7 or 14.

Continuity is reset to zero if an overdose diagnosis code appears on the same outpatient or inpatient claim.

SUD diagnoses are used to identify procedures connected to SUD diagnoses. SUD diagnoses are identified through ICD-9 codes. Procedures are defined using a combination of Healthcare Common Procedure Coding System (HCPCS) codes, Uniform Billing (UB) Revenue Codes and ICD-9/ICD-10 procedure codes.

Pharmacotherapy includes naltrexone (short or long acting), acamprosate, or disulfiram for alcohol dependence treatment and buprenorphine for opioid dependence treatment, as well HCPCS codes to identify procedures related to injecting drugs (e.g., long-acting injectable naltrexone).

A list of value sets for the measure is attached in the Excel workbook provided for question S. 2b. States may need to adapt the list of codes to include state-specific codes.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Adult Medicaid beneficiary discharges from detoxification from January 1 to December 15 of the measurement year.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

Measure data will be reported annually (12 months). To account for the 14-day time period after discharge from detoxification, the denominator period will start January 1 and end December 15 of the measurement year.

Target population meets the following conditions:

• Medicaid beneficiaries aged 18 years and older and less than 65 years with at least one detox discharge during the year January 1-December 15.

• Enrolled in Medicaid during the month of detoxification discharge and the following month.

The denominator is based on discharges, not individuals. A beneficiary may have more than one qualifying detox episode.

Detoxification is identified using a combination of HCPCS codes, UB Revenue Codes and ICD-9/ICD-10 procedure codes. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b. As with the numerator specifications, this document lists standardized specification for this measure. States will likely need to modify the specifications to include their state-specific codes.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Not applicable. The measure does not have denominator exclusions.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.)*

Not applicable.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.)*

Location of detox is used as a stratification variable in analyses. If an inpatient hospital claim had an ICD-9/ICD-10 detoxification procedure code or a UB revenue code indicating detoxification, hospital inpatient treatment is assigned as the location of detox. In addition, hospital inpatient treatment is also assigned if a non-inpatient claim contains a HCPCS code indicating hospital inpatient detox. The remaining detox location assignments are very straightforward. Whenever possible, use of the HCPCS codes to determine location is most desired as it reflects the more precise detoxification location. The other stayover treatment location is designed to capture detox location from non-inpatient claims that do not contain a HCPCS code. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

The following step are used to identify the denominator, numerator, and calculation of the measure rate:

Step 1: Identify denominator

Step 1A: Eligible population: Identify enrolled Medicaid beneficiaries ages 18-64 years who have any detoxification (withdrawal management) in inpatient hospital, residential addiction treatment program, or ambulatory detoxification (withdrawal management) discharge from January 1 to December 15 of the measurement year and are enrolled the month of detoxification and the following month. Age is calculated as of January 1 of the measurement year.

Step 1B: Overall: Among the Medicaid beneficiaries in Step 1A, identify all detoxification discharges using all inpatient, outpatient and ambulatory claims files or tables that contain HCPCS or ICD-9/ICD-10 procedure codes and UB revenue codes. If more than one detoxification in a year, treat each detoxification as a separate observation, e.g., an inpatient hospital detoxification in January and an ambulatory detoxification in July, counts as two observations.

Step 1B.1: Multiple detox claims that are within 1-2 days are combined into a single detox episode. Accordingly, sort the inpatient, outpatient and ambulatory detox discharges by Beneficiary ID and service dates to ensure the discharges from these multiple data sources are in chronological order. Then combine close-proximity episodes while retaining all clinical fields from each episode.

Step 1C: Detox location assignment: hospital inpatient, inpatient residential addiction, outpatient residential outpatient addiction, other stayover treatment and ambulatory detoxification. Use HCPCs detox procedure codes to assign detox location whenever possible; revenue center detox will map to the hospital inpatient location when the revenue codes appear on an inpatient claim or table. They will map to other stayover treatment when the revenue codes appear on a non-inpatient claim. If there is more than 1 detox location when episodes are combined, assign the location using the first claim´s location. If there is a TIE between a detox episode being identified via revenue center codes and a more specific category using HCPCs on the SAME claim, the HCPCs location prevails.

Step 2: Identify numerator

Step 2A: Overall: From the denominator in Step 1B, identify those discharges from detoxification in any setting with a qualifying continuity service within 7 or 14 days after discharge.

Step 2A.1: Identify SUD continuity services: Continuity services are assigned using clinical claims billing information (e.g., diagnosis, procedure, revenue codes). The measure includes all claims files or data tables that contain clinical fields (e.g., inpatient hospital, outpatient, other ambulatory and long-term care). SUD diagnoses can be in any position – primary or secondary – for continuity services. Since multiple claims files or tables could each contain a continuity claim, the specification calls for creating continuity variables separately within each file type or table, sorting the files or tables by beneficiary ID and service dates, then putting them together in order to assign the set of variables that are “First” to occur relative to the detox episode discharge date. Continuity services have to occur the day after discharge through day 7 or 14.

Step 2A.2: Identify pharmacotherapy which may occur in multiple files or tables. For example, one claims file or data source may contain injectables, another claims file or table data source may contain oral medications. Consequently, pharmacotherapy variables are created separately in each source, the data sources are then sorted by beneficiary ID and service dates, then multiple pharmacotherapy data sources are put together so they will be in chronological order to assign “First” variables. Pharmacotherapy services could be provided on the same day as the discharge from detox through day 7 or 14.

Step 2A.3: Co-occurring events: Continuity service flags and pharmacotherapy flags are reset to zero if an overdose diagnosis code appears on the SAME claim as the continuity service. Further, outpatient continuity is also reset to zero if an emergency department visit occurs on the same day. If an inpatient continuity claim has an emergency department visit, it is allowed to remain a continuity service.

Step 3: Calculate rate

Step 3A: Calculate the overall 7- or 14-day continuity rates by dividing the number of discharges with a qualifying continuity service (Step 2A) by the denominator (Step 1B).

Step 3B: Calculate the rates separately for each detox location by dividing the respective number of discharges by each location with a qualifying continuity service (Step 2A) by the denominator (Step 1C).

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Not applicable; this measure does not use a sample.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

Not applicable; this measure does not use a survey.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.

Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*
IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Medicaid Analytic eXtract (MAX) 2013 and 2014 eligible (EL), inpatient (IP), other services (OT), long-term care (LT) and drug (RX) files. The other services file contains facility and individual provider services data. Most notably, it may contain both residential and other stayover service claims data as claims are assigned to MAX claims file types based upon the category of service provided. The inpatient file only contains inpatient hospital, sterilization, abortion and religious non-medical health care institution claims.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Population : Regional and State

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital, Outpatient Services

If other:

S.22. 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.)*

Not applicable.

2. Validity – See attached Measure Testing Submission Form

Cont_Care_Afer_Detox_testing_attachment.pdf

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 <u>For maintenance of endorsement</u>

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 <u>For maintenance of endorsement</u>

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

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

Measure Title: Continuity of Care for Medicaid Beneficiaries after Detoxification (Detox) From Alcohol and /or Drugs

Date of Submission: 10/31/2017

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures**, section **2b3** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins).

  Contact NQF staff if more pages are needed.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

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

- **2a2. Reliability testing** [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.
- **2b1. Validity testing** [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

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

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.17***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: eligibility data from Medicaid Person Summary file	☒ other: eligibility data from Medicaid Person Summary file

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*).

Medicaid Analytic eXtract (MAX) 2013 and 2014 eligible (EL), inpatient (IP), other services (OT), long- term care (LT) and drug (RX) files. The other services file contains facility and individual provider services data. Most notably, it may contain both residential and other stayover service claims data as claims are assigned to MAX claims file types based upon the category of service provided. The inpatient file only contains inpatient hospital, sterilization, abortion and religious non-medical health care institution claims.

We used the following MAX Medicaid files to identify adult Medicaid beneficiaries with discharges from detox (denominator) and the qualifying substance use treatment services and pharmacotherapy (numerator):

Person Summary (PS): Person-level file, including Medicaid eligibility and demographic information

Inpatient (IP): Claims-level file, including information on inpatient hospital stays

Long-Term Care (LT): Claims-level file, including information on long-term care institutional stays (nursing facilities, intermediate care facilities for individuals with intellectual disabilities, psychiatric hospitals, etc.)

Other Therapy (OT): Claims-level file, including information on use of "other" services, such as home- and community-based service use Prescription Drug (RX): Information on drugs and other services provided by a pharmacy

1.3. What are the dates of the data used in testing? January-December 2014 for all states, except 2013 for Texas and Washington State.

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

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.20*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: state level	☒ other: state level

We calculated state specific SUD-5 measure variance ("noise") as a function of the measure rate at the state level for Medicaid beneficiaries within each state.

We assessed the temporal consistency (also referred to as temporal stability) of the SUD-5 measure by examining the strength of association between measure results in four quarters of the 2014 (or 2013 for Texas and Washington) measurement year for the entire sample.

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*)

We included 14 states in measure testing: Connecticut, Georgia, Iowa, Michigan, Missouri, Mississippi, New Jersey, New York, Pennsylvania, Tennessee, Texas, Vermont, Washington, and West Virginia.

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*)

Included in the testing and analyses were 47, 313 beneficiaries with at least one detoxification episode during the year. Table 1 shows the number and percent distribution by Medicaid beneficiary category, age, gender, and race/ethnicity by state.

Table 1. Testing population characteristics: Medicaid beneficiary category, age, gender, and race/ethnicity by states

Beneficiary Characteristics	Total		Connecticut		Georgia		Iowa		Michigan		Missouri		Mississippi		New Jersey	
	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)	Number of Benes with at least one Detoxification	Distribution (%)
TOTAL	47,313	100.0	2,028	100.0	1,255	100.0	336	100.0	3,315	100.0	798	100.0	618	100.0	4,734	100.0
Medicaid beneficiary category																
Aged	678	1.4	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.5	NR	0.0
Blind-disabled	13,832	29.2	267	13.2	679	54.1	53	15.8	1,027	31.0	540	67.7	348	56.3	836	17.7
Adult	31,886	67.4	1,758	86.7	569	45.3	283	84.2	2,234	67.4	248	31.1	262	42.4	3,854	81.4
Child	734	1.6	NR	0.1	NR	0.6	NR	0.0	15	0.5	NR	1.1	NR	0.8	38	0.8
Unknown	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0
Age																
18-24	5,163	10.9	114	5.6	96	7.6	12	3.6	335	10.1	60	7.5	63	10.2	439	9.3
25-44	24,417	51.6	933	46.0	634	50.5	136	40.5	1,727	52.1	429	53.8	341	55.2	2,555	54.0
45-64	17,550	37.1	981	48.4	525	41.8	188	56.0	1,215	36.7	308	38.6	214	34.6	1,734	36.6
Gender																
Male	28,547	60.3	1,381	68.1	553	44.1	241	71.7	1,727	52.1	444	55.6	266	43.0	2,934	62.0
Female	18,583	39.3	647	31.9	702	55.9	95	28.3	1,550	46.8	353	44.2	352	57.0	1,794	37.9
Unknown	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0
Race/ethnicity																
White	27,979	59.1	1,476	72.8	676	53.9	250	74.4	2,125	64.1	558	69.9	349	56.5	2,856	60.3
Black	9,449	20.0	241	11.9	268	21.4	24	7.1	862	26.0	201	25.2	215	34.8	958	20.2
American Indian/Alaskan Native	300	0.6	NR	0.2	NR	0.2	NR	2.4	29	0.9	NR	0.8	NR	0.2	NR	0.2
Asian	556	1.2	NR	0.4	NR	0.2	NR	0.3	NR	0.1	NR	0.0	NR	0.2	19	0.4
Hispanic/Latino	5,905	12.5	298	14.7	NR	0.3	NR	1.8	59	1.8	13	1.6	NR	0.6	331	7.0
Native Hawaiian/Pacific Islander	51	0.1	NR	0.0	NR	0.0	NR	0.0	NR	0.1	NR	0.0	NR	0.0	NR	0.0
Other Race/Ethnicity	201	0.4	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0
Unknown Race/Ethnicity	2,689	5.7	NR	0.0	303	24.1	47	14.0	197	5.9	19	2.4	48	7.8	554	11.7

Beneficiary Characteristics	New York		Pennsylvania		Tennessee		Texas		Vermont		Washington		West Virginia	
	Number of Benes with at least one Detoxification	Percent Distribution	Number of Benes with at least one Detoxification	Percent Distribution	Number of Benes with at least one Detoxification	Percent Distribution	Number of Benes with at least one Detoxification	Percent Distribution	Number of Benes with at least one Detoxification	Percent Distribution	Number of Benes with at least one Detoxification	Percent Distribution	Number of Benes with at least one Detoxification	Percent Distribution
TOTAL	19,473	100.0	7,322	100.0	3,203	100.0	774	100.0	1,008	100.0	1,625	100.0	824	100.0
Medicaid beneficiary category														
Aged	NR	0.0	NR	0.0	NR	0.0	NR	0.0	662	65.7	NR	0.1	NR	0.0
Blind-disabled	3,419	17.6	4,013	54.8	1,051	32.8	517	66.8	84	8.3	830	51.1	168	20.4
Adult	15,819	81.2	2,905	39.7	2,067	64.5	246	31.8	233	23.1	754	46.4	654	79.4
Child	134	0.7	362	4.9	85	2.7	NR	1.2	29	2.9	39	2.4	NR	0.0
Unknown	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0
Age														
18-24	1,402	7.2	1,343	18.3	542	16.9	91	11.8	263	26.1	342	21.0	61	7.4
25-44	9,255	47.5	4,120	56.3	1,986	62.0	382	49.4	592	58.7	839	51.6	488	59.2
45-64	8,724	44.8	1,817	24.8	675	21.1	299	38.6	153	15.2	444	27.3	273	33.1
Gender														
Male	14,459	74.3	3,361	45.9	1,109	34.6	359	46.4	545	54.1	604	37.2	564	68.4
Female	4,922	25.3	3,919	53.5	2,094	65.4	413	53.4	463	45.9	1,021	62.8	258	31.3
Unknown	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0	NR	0.0
Race/ethnicity														
White	8,395	43.1	5,448	74.4	2,598	81.1	350	45.2	950	94.2	1,146	70.5	802	97.3
Black	4,905	25.2	1,278	17.5	292	9.1	78	10.1	15	1.5	92	5.7	20	2.4
American Indian/Alaskan Native	84	0.4	13	0.2	NR	0.1	NR	0.1	NR	0.2	136	8.4	NR	0.0
Asian	495	2.5	12	0.2	NR	0.0	NR	0.5	NR	0.0	NR	0.6	NR	0.0
Hispanic/Latino	4,423	22.7	432	5.9	11	0.3	203	26.2	NR	0.5	116	7.1	NR	0.0
Native Hawaiian/Pacific Islander	36	0.2	NR	0.0	NR	0.1	NR	0.0	NR	0.0	11	0.7	NR	0.0
Other Race/Ethnicity	NR	0.0	NR	0.0	201	6.3	NR	0.0	NR	0.0	NR	0.0	NR	0.0
Unknown Race/Ethnicity	1,043	5.4	97	1.3	95	3.0	136	17.6	36	3.6	114	7.0	NR	0.0

NR=Not reported; result is based on a cell size of 10 or less

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.

Validity testing sample: We restricted the analytic sample for validity testing to individuals enrolled in Medicaid in the month before the detox episode and 90 days after the detox episode concluded to allow time to track the incidence of poor outcomes. Moreover, we excluded beneficiaries from convergent validity testing if they had the poor outcome (overdose treatment OR detox readmission) within 14 days of detox discharge.

Reliability testing sample: For both 14- and 7-day continuity, we calculated the signal-to-noise analysis and temporal consistency across four quarters of the measurement year within each state.

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

As described in section 1.6, we collected information on the following variables using data extracted from Medicaid Analytic eXtract (MAX) 2013 and 2014 files: Medicaid eligibility category, age, gender, and race/ethnicity. This measure is based on a process that should be carried out for all beneficiaries (except those excluded), so no adjustment for patient mix is necessary. We did collect information about these variables and assessed disparities in performance rate for each group.

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*)

Signal-to-noise reliability. The signal-to-noise (SNR) statistic, R (ranging from 0 to 1), summarizes the proportion of the variation between entity scores that is due to real differences in underlying entity characteristics (such as differences in population demographics or medical care) as opposed to background-level or random variation (for example, due to measurement or sampling error). If R=0, there is no variation on the measure across entities, and all observed variation is due to sampling variation. In this case, the measure is not useful to distinguish between entities with respect to healthcare quality. Conversely, if R=1, all entity scores are free of sampling error, and all variation represents real differences between entities in the measure result.

We estimated SNR reliability for the SUD-5 measure using a beta-binomial model, which is suitable for binary pass/fail rate measures (Adams, 2009). For SUD-5, the pass/fail designation is defined as having or not having an eligible follow-up visit within a specified time frame (7 days and 14 days) after eligible discharge from a detoxification episode (an inpatient hospital, residential addiction program, or ambulatory detoxification). The beta-binomial model assumes the entity SNR score is a binomial random variable conditional on the entity's true value, which comes from the beta distribution (ranging from 0 t o1). We calculated SNR reliability in three steps (Adams, 2009, 2014):

First, we calculated state specific SUD-5 measure variance ("noise") as a function of the measure passing rate at the state level, $\hat{p}^1$ and the sample size, *n*:

$$\sigma^2_{within} = \frac{\hat{p}(1-\hat{p})}{n} \quad (1);$$

Second, we used version 2.2 of the BETABIN SAS macro written by Wakeling (n/d) to fit the beta- binomial model to the SUD-5 dataset (Wakeling, n/d) . The macro produced the estimated average pass rate across all providers, as well as the Alpha (α) and Beta (β) parameters that describe the shape of the fitted beta-binomial distribution. We calculated the "signal" (between-state variation of the SUD-5 measure) using these parameters, as follows:

$$\sigma^2_{between} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \quad (2);$$

Third, we calculated the SNR reliability as the ratio of the between-level variance and the total variance (i.e., the sum of the between-level and within-level variances) of the SUD-5 measure rate:

$$SNR = \frac{\sigma^2_{between}}{\sigma^2_{between}+\sigma^2_{within}}; \quad (3);$$

We calculated reliability of the SUD-5 measure using two alternative definitions of continuity treatment services after detoxification discharges set at 7 and 14 days as stated in the specifications.

Temporal consistency. We assessed the temporal consistency (also referred to as temporal stability) of the SUD-5 measure by examining the strength of association between the state-level measure results in four quarters of the 2014 measurement year. Specifically, we first aggregated the SUD-5 measure result for each state within each quarter in 2014, and then calculated Spearman's rank-order correlation coefficient (ranging from -1 to +1) between the state-level measure results in consecutive quarters (i.e. 2014 Q1 vs Q2, Q2 vs Q3, and Q3 vs Q4). High positive value indicates a strong tendency for the paired measure ranks to move together, whereas a negative value indicates that the paired measure ranks move in opposite directions.

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*)

Signal-to-noise reliability. Table 2 summarizes the mean and range of the SNR statistic for SUD-5, which was computed separately for each of the 14 states in the sample by each of the two definitions of continuity threshold (7-day and 14-days; Table 3). Note that the threshold definition

---

[1] Beneficiaries who had an eligible follow-up visit within a specified timeframe/Eligible beneficiaries discharged from detox

only affects the SUD-5 measure numerator (eligible follow-up visit within a specified continuity timeframe) but does not affect the measure denominator (eligible discharge from a detoxification episode). Generally, we observed smaller numerator counts using 7-day continuity threshold compare to the 14-day continuity threshold.[2]

The SUD-5 was highly reliable in distinguishing performance between States using both 7- and 14-days continuity threshold, with the average reliability score of 0.98 across states and a range from 0.98 to 0.99.

Note that high reliability is not indicative of high quality of healthcare, but rather indicates that the SUD- 5 measure can be used to distinguish between entities with respect to healthcare quality. The high reliability for the measure at the state level is likely driven by the adequate sample sizes and low "noise" variance within the States. The figure below demonstrates the relationship between the number of discharges from detox at the state level (the SUD-5 measure denominator) and the resulting SNR statistic.

Table 2. Signal-to-noise reliability for the SUD-5 measure (n=14 states)

SUD-5 Continuity Threshold	Average reliability score	Range of reliability scores
7-day	0.99	(0.99-0.99)
14-day	0.99	(0.98-0.99)

Notes: Data from 14 states were included in the analysis. Based on analysis of 2014 (2013 for Texas and Washington State) MAX PS, IP, LT, OT, and RX files.For both 7- and 14-day continuity

thresholds the signal-to-noise coefficients for GA, VT, WA, CT, TN, MI, NJ, PA and NY were truncated to 0.99 rather than rounded to 1.00 to reflect the uncertainty in the estimates.

[2] For both 7- and 14-day continuity thresholds the average number of eligible discharges from detox per State for the SUD-5 measure was 4,837 (ranging from 379 to 32,744). The average number of episodes with continuity treatment within 7 days after discharge per State was 1.392 (ranging from 78 to 6149). The average number of episodes with continuity treatment within 14 days after discharge per State was 1,765 (ranging from 107 to 8,940). These statistics are based on all eligible denominator and numerator counts in the SUD-5 dataset including those with less than 11 observations in the denominator or numerator.

Table 3. SUD-5 Measure rate and signal-to-noise reliability, by State

		7-Day Continuity Threshold			14-Day Continuity Threshold		
State abbreviation	# of eligible discharges from detox	# of discharges with continuity treatment	Mean SUD-5 rate	Signal- to-noise reliability	# of discharges with continuity treatment	Mean SUD-5 rate	Signal-to-noise reliability
IA	379	78	0.21	0.99	107	0.29	0.98
MS	884	239	0.27	0.99	303	0.34	0.99
TX	929	304	0.33	0.99	367	0.40	0.99
WV	985	380	0.39	0.99	448	0.46	0.99
MO	997	289	0.29	0.99	365	0.37	0.99
VT	1,160	944	0.81	0.99	979	0.85	0.99
GA	1,571	287	0.18	0.99	409	0.26	0.99
WA	2,058	456	0.22	0.99	603	0.29	0.99
CT	2,799	858	0.31	0.99	1,162	0.42	0.99
MI	3,760	2,368	0.63	0.99	2,537	0.68	0.99
TN	3,911	1,288	0.33	0.99	1,542	0.39	0.99
NJ	6,068	959	0.16	0.99	1,405	0.23	0.99
PA	9,474	4,882	0.52	0.99	5,544	0.59	0.99
NY	32,744	6,149	0.19	0.99	8,940	0.27	0.99

Notes: The signal-to-noise coefficients for VT, GA, WA, CT, MI, TN, NJ, PA and NY were truncated to 0.99 rather than rounded to 1.00 to reflect the uncertainty in the estimates.

Overall, using the 7-day continuity threshold we observed marginally higher signal to-noise-reliability for 13 out of 14 states (except VT for which the reliability decreased by 0.03 percentage points). On average, the 7-day continuity threshold average reliability was 0.1 percentage point (PP) higher (with the change in reliability ranging from -0.4 PP +4.7 PP) compared to the 14-day threshold.

The small increase in the SUD-5 measure reliability using the 7-day continuity threshold can be explained by examining three key drivers of reliability: "signal," "noise" and denominator sample size. First, using the 7-day continuity threshold results in somewhat larger between-state variance or stronger "signal" compared to the 14-day threshold (4 and 10 PP difference)l. Secondly, with the 7-day continuity threshold we observed smaller within-state variances (or weaker "noise") for 11 out of 14 states. The noise variance was on average 9.9 PP weaker. Since

denominator size remained the same for both definitions of continuity threshold, consistently stronger “signal” with generally weaker “noise” mostly resulted in higher reliability of the SUD-5 data with the 7-day continuity threshold.

Temporal consistency. Table 4 provides the measures of temporal consistency (Spearman rank correlation) across four quarters of the 2014 measurement year for the SUD-5 measure. Our results indicate very high (at or above 0.90) temporal stability of the SUD-5 measure over time.

Table 4. Temporal consistency of SUD-5 in the measurement year

		Spearman Rank Correlations		
SUD-5 Continuity Threshold	**Average across 4 quarters**	**Q1 vs. Q2**	**Q2 vs. Q3**	**Q3 vs. Q4**
7 Day	0.93	0.94	0.94	0.92
14 Days	0.93	0.96	0.92	0.92

Notes: All correlation coefficients are statistically significant at p<0.001; each pairwise correlation included only those States that had data during both quarters analyzed.

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?*)

SUD-5 is rated high for scientific acceptability, based on reliability testing results. Specifically, the excellent SNR indicated that the SUD-5 measure can discern the underlying performance between states within high precision. High temporal consistency showed that the performance of state-level SUD-5 rates were consistent over time.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ Performance measure score

☒ **Empirical validity testing**

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

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

We conducted validity testing at the performance score level for both 14- and 7- day continuity.

Convergent validity. To assess the convergent validity of the SUD-5 measure, we examined the association between presence/absence of continuity of care (that is, the underlying construct of the measure) - defined as having a follow-up visit within 7 or 14 days after discharge from detox—and presence/absence of a subsequent overdose treatment or detox readmission). We hypothesized that there would be fewer overdoses or detox readmissions between days 15 and 90 (for the analysis of 14- day continuity) or between days and 8 and 90 (for the analysis of 7-day continuity) after the detox among beneficiaries with continuity of care compared to those without. We used inverse probability

weighting and doubly robust regression (Imbens & Wooldridge, 2009) to examine the association and controlled for potential confounders, including age group, gender, race/ethnicity, focus of detox, location of detox, blind/disabled status, the use of pharmacotherapy for SUD in the 30 days before the detox treatment of interest, and the use of behavioral and/or physical health services (in an inpatient, outpatient, or emergency department setting) in the 30 days before the detox treatment of interest.

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

Convergent validity. The odds of subsequent overdose treatment or readmissions between days 15–90 among those with continuity of care within 14 days were 0.917 (with 95% confidence interval (0.863, 0.976)) as much as those of individuals without continuity of care, translating into an absolute risk reduction of 1.4 percent and a number needed to treat (NNT) of 71, which was statistically significant ($p < 0.01$). Similar results hold when looking into the outcome with a 7-day threshold. The odds of subsequent overdose treatment or readmissions between days 8–90 among those with continuity of care were 0.834 (with 95% confidence interval (0.783, 0.889)) as much as those of individuals without continuity of care, translating into an absolute risk reduction of 2.1 percent and an NNT of 48, which was statistically significant ($p < 0.01$).

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

SUD-5 is also rated high for scientific acceptability, based on validity results. The convergent validity of SUD-5 was excellent, with a lower odds (e.g. 8.3% lower for those with continuity of care within 14 days) of readmission to detox or overdose treatment among detox episodes with continuity.

2b2. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section* 2b3

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

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section* 2b4.**

Not applicable - Not an intermediate or health outcome, or PRO-PM, or resource use measure.

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

☐ **No risk adjustment or stratification**

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

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

☐ **Other,** Click here to enter description

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

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

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors*

identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

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

2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*): 2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves: 2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

We compared performance across state-level continuity rates to understand any variation in performance. We examined the distribution of the measure (for example, mean, median, minimum, 25th percentile, 75th percentile, and maximum) across states. In addition, we calculated the 95% confidence interval of the continuity rates for each state using a z-distribution for proportion. Then we compared each state's confidence interval to the overall measure rate that uses all beneficiaries across states. States measure rates significantly lower than the overall rate indicate an evidence of less-than- optimal performance, hence room for improvement.

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

For both 14-day and 7-day continuity rates, we found a wide range with meaningful variation. The 7-day continuity rate ranges from 15.80 percent to 81.38 percent with a median of 29.82 percent, and a mean of 34.46 percent (Table 5). The 14-day continuity rate ranges from 23.15 percent to 84.40 percent with a median of 38.02 percent, and a mean of 41.52 percent.

Table 5. Distribution of the SUD-5 Measure Rate

	Distribution of the Measure Rate (%)					
	Minimum	**25th Percentile**	**50th Percentile**	**75th Percentile**	**Maximum**	**Mean**
7-day continuity	15.80	20.97	29.82	37.17	81.38	34.46
14-day continuity	23.15	28.50	38.02	44.49	84.40	41.52

Note: Data from 14 states are included in the analyses.

For both 7-day and 14-day continuity, the rates are greater than 50 percent in Vermont, Michigan, and Pennsylvania (Table 6 and Table 7). New Jersey had the smallest measure rate.

For the 14-day continuity measure rate, the z-test for proportion indicates that 6 states have a measure rate significantly[3] greater than the overall measure rate, 5 states have a measure rate significantly lower than the overall measure rate, and the remaining 3 states had measure rates which were indistinguishable from the overall measure rate (Table 6).

Table 6. State-level SUD5 Measure Rate (14-day Continuity)

State	**Number of Detoxification Episodes**	**Number of Detoxification Episodes with Continuity**	**Percentage of Detoxification Episodes with Continuity**	**95 Confidence Interval**
Total	**67,719**	**24,711**	**36.49%**	
Connecticut*	2,799	1,162	41.51%	(39.69, 43.34)
Georgia†	1,571	409	26.03%	(23.86, 28.20)
Iowa†	379	107	28.23%	(23.70, 32.76)
Michigan*	3,760	2,537	67.47%	(65.98, 68.97)
Missouri	997	365	36.61%	(33.62, 39.60)
Mississippi	884	303	34.28%	(31.15, 37.40)
New Jersey†	6,068	1,405	23.15%	(22.09, 24.22)
New York†	32,744	8,940	27.30%	(26.82, 27.79)
Pennsylvania*	9,474	5,544	58.52%	(57.53, 59.51)
Tennessee*	3,911	1,542	39.43%	(37.90, 40.96)
Texas	929	367	39.50%	(36.36, 42.65)
Vermont*	1,160	979	84.40%	(82.31, 86.48)
Washington†	2,058	603	29.30%	(27.33, 31.27)
West Virginia*	985	448	45.48%	(42.37, 48.59)

[3] Statistical significance is defined throughout as significant at the 0.05 level.

Source: Based on analysis of 2014 (2013 for Texas and Washington State) MAX PS, IP, LT, OT, and RX files.

Note: * Significantly greater than the total measure rate at the .05 level. † Significantly less than the total measure rate at the .05 level.

For the 7-day continuity, the z-test for proportion indicates that 7 states have a measure rate significantly greater than the overall measure rate, 5 states have a measures rate significantly lower than the overall measure rate, and the remaining 2 states have measure rates which were indistinguishable from the overall measure rate (Table 7).

Table 7. State-level SUD5 Measure Rate (7-day Continuity)

State	Number of Detoxification Episodes	Number of Detoxification Episodes with Continuity	Percentage of Detoxification Episodes with Continuity	% Confidence Interval
Total	**67,719**	**19,481**	**28.77%**	
Connecticut*	2,799	858	30.65%	(28.95, 32.36)
Georgia†	1,571	287	18.27%	(16.36, 20.18)
Iowa†	379	78	20.58%	(16.51, 24.65)
Michigan*	3,760	2,368	62.98%	(61.44, 64.52)
Missouri	997	289	28.99%	(26.17, 31.80)
Mississippi	884	239	27.04%	(24.11, 29.96)
New Jersey†	6,068	959	15.80%	(14.89, 16.72)
New York†	32,744	6,149	18.78%	(18.36, 19.20)
Pennsylvania*	9,474	4,882	51.53%	(50.52, 52.54)
Tennessee*	3,911	1,288	32.93%	(31.46, 34.41)
Texas*	929	304	32.72%	(29.71, 35.74)
Vermont*	1,160	944	81.38%	(79.14, 83.62)
Washington†	2,058	456	22.16%	(20.36, 23.95)
West Virginia*	985	380	38.58%	(35.54, 41.62)

Source: Based on analysis of 2014 (2013 for Texas and Washington State) MAX PS, IP, LT, OT, and RX files.

* Significantly greater than the total measure rate at the .05 level.

† Significantly less than the total measure rate at the .05 level.

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

SUD-5 is rated high for validity, based on statistically significant and meaningful differences. The measure results suggest variation in performance and room for improvement in continuity of care after detoxification. For 14-day and 7- day continuity measures, five states had a continuity rate significantly below the overall total rate (Table 6 and Table 7). For the 14-day measure, three states had performance not distinguishable from the average performance; and for the 7-day measure, two states had performance not distinguishable from the average.

It is important to note that interpretation of the results should be tempered by the fact that only 14 states are included in the total; the total continuity rate for the entire nation could be different. The total is also weighted more heavily toward larger states, and states differ in terms of which detox and continuity services Medicaid covers. In terms of room for improvement, even states that are not statistically different from or even above the overall total of 36.5 percent have room for improvement. Achieving continuity of treatment after detoxification should be a goal for all clients, and only two states reach a rate greater than 60 percent for 7-day or 14-day continuity.

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

***If only one set of specifications, this section can be skipped*. Not applicable; only one set of specifications.**

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

We assessed the extent of missing data was using the MAX validation and anomaly tables. These tables are available online at:

- MAX validation tables: https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer- Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports.html?DLSort=0&DLEntries=10&DLPage=1&DLSortDir=ascending.
- MAX anomaly tables: https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer- Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html.

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

SUD-5 is a claims-based measure that relies on National Drug Code (NDC) in the RX file and procedure and diagnosis codes in the IP, LT, and OT files. Missing data is not a concern for many of the MAX data elements used to construct the SUD-5 measure in the study states.

- The service ending dates in the IP, OT, and LT files are non-missing because claims are assigned to yearly files using ending date; as such, a claim must have a non-missing ending date to be included in the MAX data. Similarly, prescription fill dates in the RX files are non-missing because RX claims are assigned to the yearly RX file using prescription fill date. Service beginning dates are infrequently missing.
- We found NDC to be non-missing in RX files.
- The SUD-5 specification utilizes secondary (and beyond) procedure and diagnosis codes; however, in the

validation and anomaly tables, missing information is documented only for the primary diagnosis code and "a" procedure code. The absence of secondary primary and procedure codes may reflect missing data or may reflect the beneficiary's true clinical.

- Among the study states, the primary diagnosis code is mostly non-missing in the IP and LT files (Table 8). Missingness of primary diagnosis code in the OT file and procedure code in the IP and OT files varies by study state. For example, the percent of OT claims with a primary diagnosis code ranged from 57.5 percent in Washington to 98.8 percent in Vermont (Table 8). In most states, most claims had a procedure code in the OT file. Procedure code in the IP file had higher rates of missingness in each state than in the OT file. Missing procedure and diagnosis codes may result in mistakenly excluding beneficiaries from the denominator or numerator, increasing the risk of over- or under-estimating the measure rate.

In New York and New Jersey, we found that the states were using state-specific codes for methadone treatment claims, which would not be currently captured by the measure specifications. In addition, New York frequently uses state-specific procedure codes. In the measure submission form, we advise measure implementers to include the relevant state-specific codes in the measure specification and calculation. Accounting for state-specific codes will improve the accuracy of measures calculated by states.

Table 8. Percent of IP, LT, or OT file with primary diagnosis code, procedure code, or place of service

State	Percent with primary diagnosis code			Percent with procedure code		Percent with place of service
	IP	LT	OT	IP	OT	OT
Connecticut	100.0	100.0	88.8	58.4	91.3	92.3
Georgia	100.0	100.0	95.7	60.5	96.3	88.3
Iowa	100.0	94.4	89.1	65.9	100.0	87.7
Michigan	100.0	100.0	80.3	68.3	99.7	99.9
Mississippi	100.0	100.0	83.9	31.8	99.1	79.2
Missouri	100.0	100.0	97.5	42.9	100.0	93.1
New Jersey	100.0	100.0	97.4	69.2	96.7	90.4
New York	100.0	100.0	97.4	74.8	99.2	87.6
Pennsylvania	100.0	100.0	97.3	67.4	100.0	75.7
Tennessee	0.0	100.0	58.9	0.0	100.0	100.0
Texas	100.0	98.9	65.5	66.5	83.0	67.2
Vermont	100.0	100.0	98.8	58.3	91.6	92.9
Washington	100.0	100.0	57.5	61.3	99.8	88.9
West Virginia	100.0	100.0	90.7	59.7	98.9	96.4

Source: MAX anomaly tables. Available at the following URL: https://www.cms.gov/Research-Statistics-Data-and- Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html.

Note: Numbers are from 2013 for all study states except Texas; the most recent numbers available for Texas are from 2012.

We used two additional variables to create the measure – UB-92 revenue codes and place of service from the OT file. The percent of OT claims with a valid place of service ranges from 75.7 percent in Pennsylvania to 100 percent in Tennessee (Table 8).

To calculate the SUD-5 measure generally and for specific subgroups, we also use data elements from the MAX PS file, including race, sex, zip code, age (calculated using date of birth), information about prepaid plans, and eligibility information. Sex and date of birth are rarely missing (Table 9). Nearly all enrollees have a valid 5-digit zip code. Race, however, is missing for a substantial portion of enrollees in some states (for example, 43.8 percent of enrollees in Iowa), so examination of SUD-5 by race subgroup will exclude beneficiaries who are missing race data. Information about prepaid plans are generally non- missing. Over 95 percent of MAX claims have corresponding Medicaid eligibility information (Table 10).

Table 9: Percent of Medicaid enrollees with missing date of birth, sex, or race

State	Year	Percent of Enrollees Missing Date of Birth	Percent of Enrollees with Missing Sex	Percent of Enrollees with Missing Race
Connecticut	2012	0.0	0.0	0.0
Georgia	2013	0.0	0.0	11.1
Iowa	2013	0.0	0.0	43.8
Michigan	2012	0.0	0.0	10.7
Mississippi	2013	0.0	0.0	6.1
Missouri	2012	0.0	0.0	4.1
New Jersey	2012	0.0	0.0	28.0
New York	2013	1.3	1.0	7.7
Pennsylvania	2013	0.0	0.0	12.3
Tennessee	2013	0.0	0.0	10.9
Texas	2012	0.0	0.0	60.5
Vermont	2013	0.0	0.0	26.2
Washington	2013	0.0	0.0	31.2
West Virginia	2013	0.0	0.0	1.5

Source: MAX anomaly tables. Available at the following URL: https://www.cms.gov/Research-Statistics-Data-and- Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html.

Table 10: Percent of claims missing corresponding Medicaid eligibility information

State	Year	% with Claims and Missing Medicaid Eligibility (Excludes S-CHIP Only)	IP: % Missing Eligibility and > $0 Paid (Excludes S-CHIP Only)	LT: % Missing Eligibility and > $0 Paid (Excludes S-CHIP Only)	OT: % Missing Eligibility and > $0 Paid (Excludes S-CHIP Only)
Connecticut	2013	0.27	0.22	0.07	0.18
Georgia	2013	0.96	0.12	0.02	0.17
	2014	0.85	0.07	0.01	0.16
Iowa	2013	0.17	0.14	0.04	0.01
	2014	0.08	0.04	0.02	0.00
Michigan	2013	4.08	1.59	0.41	0.38
	2014	1.50	0.94	0.46	0.10
Missouri	2013	2.03	0.14	0.01	0.97
	2014	0.35	0.21	0.02	0.07
Mississippi	2013	0.11	0.54	0.02	0.04
	2014	0.23	0.28	0.02	0.10
New Jersey	2013	0.55	0.20	0.42	0.21
	2014	0.55	0.24	0.33	0.21
New York	2013	0.07	0.23	0.21	0.00
Pennsylvania	2013	2.82	1.01	0.47	0.08
	2014	3.77	0.94	0.16	0.31
	2014	0.01	0.00	0.00	0.00
Tennessee	2013	0.39	0.00	0.00	0.03
Texas	2012	12.23	0.28	0.01	1.21
	2014	0.56	0.00	0.00	0.03
Vermont	2013	0.53	0.93	0.44	0.17
	2014	0.22	0.41	0.29	0.04
Washington	2013	1.45	0.17	0.01	0.16
West Virginia	2013	3.60	0.14	0.01	0.19
	2014	0.09	0.05	0.02	0.01

Source: MAX validation tables. Available at the following URL: https://www.cms.gov/Research-Statistics-Data-and- Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports.html?DLSort=0&DLEntries=10&DLPage=1&DLSortDir=ascending.

Note: Missing information is available for all but one of the study states in 2013. The most recent available information for Texas is for 2012. We have also provided 2014 information where available in the study states.

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

Given the relatively small amount of missing information, we don't believe there is systematic bias. In addition, states implementing the measure will likely have even less missing data than reported here because they will be able to account for their state-specific codes when constructing the measure.

References

Adams, J. L. (2009). The Reliability of Provider Profiling. A Tutorial. http://www.rand.org/pubs/technical_reports/TR653.html

Adams, J. L. (2014). Reliability-Testing Concepts. National Quality Forum presentation. Retrieved from www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=74717

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature, American Economic Association, 47*(1), 5-86.

Wakeling, I. (n/d). SAS Macro for fitting Beta-Binomial models. Retrieved from http://www.qistats.co.uk/BetaBinomial.html

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.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

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*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic claims

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

Not applicable.

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Not applicable.

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)*.

Not applicable, no fees or licensing are currently required.

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.

Specific Plan for Use	Current Use (for current use provide URL)
Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

Not applicable; the measure is under initial endorsement review and is not currently used in an accountability program.

4a1.2. 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?*)

Not applicable; this is a new measure under initial endorsement review and is not currently used in an accountability program.

4a1.3. 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.*)

CMS is developing measures to improve the quality of care of the following Medicaid populations served by CMS's Innovation Accelerator Program:

• People eligible for both Medicare and Medicaid, or "Dual-eligible beneficiaries"

• People receiving long-term services and supports (LTSS) through managed care organizations

• People with substance use disorders; beneficiaries with complex care needs and high costs; beneficiaries with physical and mental health needs; or Medicaid beneficiaries who receive LTSS in the community

This measure is intended for voluntary use by states to monitor and improve the quality of care provided for Medicaid beneficiaries with substance use disorders. States may choose to begin implementing the measures based on their programmatic needs.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Not applicable.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Not applicable.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Not applicable.

4a2.2.2. Summarize the feedback obtained from those being measured.

Not applicable.

4a2.2.3. Summarize the feedback obtained from other users

Not applicable.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Not applicable.

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

This measure is being considered for initial endorsement. Adoption of this performance measure has the potential to improve the quality of care for Medicaid beneficiaries, who have a SUD, after they are discharged from detox for alcohol and/or drugs. Currently the rate is 36.5 percent continuity within 14 days and 28.8 percent within 7 days, so there is an opportunity for improvement. The process of selection and testing of this measure was guided by the priorities outlined in the SAMHSA National Behavioral Health Quality Framework (NBHQF). The NBHQF goals reflect an effort to harmonize and prioritize measures that reflect the core principles of SAMHSA, as well as support the CMS National Quality Strategy. Specifically, this measure will encourage detox facilities to monitor the rate at which patients have follow-up treatment services to achieve continuity of care, and to take steps to put interventions in place to improve the rate with which their patients receive additional services after leaving detox. Continuity of care has been shown to reduce detox readmissions (Carrier et al., 2011; Ford and Zarate, 2010; Lee et al., 2014; Mark et al., 2006), substance use (McCusker et al., 1995; McLellan et al., 2005), and criminal justice involvement (Ford and Zarate, 2010).In addition, there is some evidence of improved employment status (Ford and Zarate, 2010).

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

Not applicable. This measure has not been implemented yet. There were no unexpected findings identified during testing of this measure.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

Not applicable. This measure has not been implemented yet. There were no unexpected findings identified benefits during testing of this measure.

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.

Yes

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

0004 : Initiation and Engagement of Alcohol and Other Drug Dependence Treatment (IET)

2605 : Follow-Up After Emergency Department Visit for Mental Illness or Alcohol and Other Drug Dependence

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

- 7-day Follow-up after Withdrawal Management; American Society of Addiction Medicine
- Continuity of Care after Detoxification; Washington Circle
- Initiation after Outpatient/Intensive Outpatient; Washington Circle

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

No

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

Follow-up time period: NQF 2605 examines follow-up care 7 days and 30 days after discharge. Our proposed measure (#3312) examines follow-up care 7 days and 14 days after discharge. The 14 day follow-up time period aligns with NQF 0004 and the non-NQF endorsed Continuity of Care After Detoxification measure developed by the Washington Circle, and reflects the input of some public commenters that adults should receive some type of care within two weeks of discharge from detoxification.

Follow-up location: NQF 2605 includes outpatient visits, intensive outpatient visits or partial hospitalizations. Our proposed measure (#3312) includes the same locations as 2605 in addition to pharmacotherapy on day of discharge, residential treatment, and long-term care. These additional follow-up services are all valid and appropriate treatments

after detoxification. If we were to exclude them from the measure specifications, measure implementers would undercount the number of beneficiaries receiving continuity of care after detoxification.

Diagnoses: NQF 2605 requires a primary diagnosis of alcohol and other drug dependence (AOD) for the follow-up service. Our proposed measure (#3312) requires a primary or secondary diagnosis of AOD. We allow a primary or secondary AOD diagnosis to address potential inaccuracies in how AOD diagnoses are coded. For example, some providers may be concerned about the stigma associated with an AOD diagnosis and therefore code it as a secondary diagnosis. Also, for adults with co-occurring mental health and AOD disorders, the assignment of primary and secondary diagnoses can be challenging and sometimes arbitrary.

The differences in follow-up time period, location and diagnoses between NQF 2605 and our proposed measure (3312) do not impact the measure's interpretability in which a higher rate is indicative of better quality. Both measures rely on administrative data. The differences in measure specifications between 2605 and 3312 are minor and expected to have minimal impact on 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.)
Not applicable. There are no other NQF-endorsed measures that conceptually address the same measure focus and same target population.

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.

No appendix Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services

Co.2 Point of Contact: Roxanne, Dupert-Frank, Roxanne.Dupert-Frank@cms.hhs.gov, 410-786-9667-

Co.3 Measure Developer if different from Measure Steward: Mathematica Policy Research

Co.4 Point of Contact: Henry, Iryes, hireys@mathematica-mpr.com, 202-554-7536-

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.

Consumer Representative 1
-Carol McDaid
-Capitol Decisions, Inc
Consumer Representative 2
-Janice Tufte
-Patient-Centered Outcomes Research Institute (PCORI) ambassador
-PCORI
Consumer Representative 3
-Kayte Thomas
-PCORI ambassador
- PCORI
State Official 1
-Joe Parks
-Missouri HealthNet Division (Medicaid)
State Official 2
-David Mancuso
-Washington State Department of Social and Health Services
State Official 3
-Roxanne Kennedy
-New Jersey Division of Mental Health and Addiction Services
Health Plan Representative 1
-Alonzo White
-Aetna Medicaid
Health Plan Representative 2
-Deb Kilstein
- Association for Community Affiliated Plans
Health Plan Representative 3
-Jim Thatcher
- Massachusetts Behavioral Health Partnership, Beacon Health Options
Provider Organization Representive 1
-Daniel Bruns
-Health Psychology Associates
Provider Organization Representive 2
-Aaron Garman
- Coal Country (ND) Community Health Center (and American Academy of Family Practice Comm. on Quality & Practice)
Provider Organization Representive 3
-Annette DuBard
-Community Care of North Carolina
Subject Matter Expert/Researcher 1
- Andrew Bindman
-University of California San Francisco (incoming AHRQ director)

Subject Matter Expert/Researcher 2

-Mady Chalk

-Treatment Research Institute

Subject Matter Expert/Researcher 3

- Kimberly Hepner

-RAND Corporation

Subject Matter Expert/Researcher 4

- Benjamin Miller

-University of Colorado School of Public Health

Subject Matter Expert/Researcher 5

- Alex Sox-Harris

-Department of Veterans Affairs

Federal Agency Official 1

- Deb Potter

-Office of the Assistant Secretary for Planning and Evaluation

Federal Agency Official 2

- Laura Jacobus-Kantor

- Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality

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? Specifications for this measure will be reviewed and updated annually.

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement: Limited proprietary coding is contained in the Measure specifications for user convenience. Users of proprietary code sets should obtain all necessary licenses from the owners of the code sets. Mathematica disclaims all liability for use or accuracy of any CPT or other codes contained in the specifications.

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments:

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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3313

Measure Title: Follow-Up Care for Adult Medicaid Beneficiaries Who are Newly Prescribed an Antipsychotic Medication

Measure Steward: Centers for Medicare and Medicaid Services (CMS)

Brief Description of Measure: Percentage of new antipsychotic prescriptions for Medicaid beneficiaries age 18 years and older who have completed a follow-up visit with a provider with prescribing authority within four weeks (28 days) of prescription of an antipsychotic medication.

Developer Rationale: Among individuals with serious mental illness, physical health problems such as cardiovascular disease, metabolic disorders, and infectious disease are more prevalent compared to the general population. Antipsychotic medications can exacerbate existing physical problems as well as increase a patient's risk for developing new health concerns such as metabolic complications. Timely follow-up with a provider following the prescription of antipsychotic medications is an essential first step to ensure that physical impacts of antipsychotic medications are identified and addressed early. Early follow up is also critical to monitor for treatment effectiveness and modify dosage as necessary, as well as to identify and address any barriers to treatment adherence. By proactively following up with patients who are prescribed antipsychotic medications, providers can identify problems early in the course of treatment and minimize potential harms associated with use of those medications. Regardless of the care setting in which a patient is being treated, comprehensive assessment of both physical and mental health factors is an essential aspect of treatment with antipsychotic medications.

Numerator Statement: Antipsychotic prescriptions from the denominator prescribed to a beneficiary who completed a follow-up visit with a provider with prescribing authority within four weeks of prescription of an antipsychotic medication.

Denominator Statement: New antipsychotic prescriptions for Medicaid beneficiaries age 18 years and older.

Denominator Exclusions: • Medicaid beneficiaries with an acute inpatient admission during the four-week follow-up period after prescription of an antipsychotic medication

• Patients who expired within four weeks of new prescription date.

Measure Type: Process

Data Source: Claims

Level of Analysis: Population : Regional and State

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* 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. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

The developer provides a logic model outlining the importance of timely follow-up post prescription of antipsychotic medications to monitor treatment effectiveness and to monitor physical impacts of medication.

The developer provides a systematic review of the evidence including four guidelines:

- American Diabetes Association (ADA), American Psychiatric Association (APA), American Association of Clinical Endocrinologists (AACE), North American Association for the Study of Obesity (NAASO). "Consensus development conference on antipsychotic drugs and obesity and diabetes." 2004. **No Grade Provided.**
- American Psychiatric Association; Work Group on Schizophrenia. "Practice guideline for the treatment of patients with schizophrenia, second edition." 2004. **Grade II (moderate clinical confidence) assigned to the recommendation. No Grade assigned to evidence**.
- University of South Florida College of Behavioral and Community Sciences. "2015 Florida Best Practice Psychotherapeutic Medication Guidelines for Adults." December 2015. **No Grade Provided.**
- University of South Florida College of Behavioral and Community Sciences. "A Summary for Monitoring Physical Health and Side-Effects of Psychiatric Medications in the Severely Mentally Ill Population." March 2014. **No Grade Provided.**

The developer provides additional evidence from literature review and clinical advisory workgroup.

Exception to evidence

N/A

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: moderate; Consistency: high (Box 5) → Moderate (Box 5b) → Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

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

Maintenance measures – increased emphasis on gap and variation

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

The developer provides rationale for the measure based on evidence summary. Measure seeks to improve quality of care for individuals who are prescribed antipsychotic medications by monitoring follow-up care.
Performance score analysis is provided based on a year of Medicaid Analytic eXtract (MAX) and Medicare claims data from 15 states: Arkansas, Connecticut, Georgia, Iowa, Michigan, Mississippi, Missouri, New Jersey, New York, Pennsylvania, South Dakota, Tennessee, Vermont, West Virginia, and Wyoming.
The overall measure performance across all states was 48.8 percent (130,785 follow-ups to 267,831 new prescriptions):

- Mean: 50.2%
- Std. Deviation: 5.3%
- Min: 44.5%
- 25th Percentile: 46%
- Median: 47.9%
- 75th Percentile: 54.3%
- Max: 60.1%
- Interquartile Range: 8.2%

Disparities
The measure performance was stratified for disparities by age, gender, race, disability status, and beneficiary category (Medicaid-only or dually eligible for Medicaid and Medicare).
Additional literature cites racial and ethnic differences in the way patients seek treatment for mental illness as well as the way mental illness is managed.
Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence
Comments:
** Four practice guidelines reviewed suggest that timely follow-up is needed for patients initiated on antipsychotic treatment. The guidelines specify that certain health parameters should be addressed including weight, blood pressure, pulse. However, the specifications of the measure address that a visit with a prescriber happened--not necessarily what is addressed during that follow-up visit.
** The developer presents a logic model, a systematic review of the literature (guidelines), consulted with stakeholders and clinical experts along with an evaluation of existing performance measures and an environmental scan. In the guidelines reviewed, details of consistency of reviewed evidence were not presented.
The data applies directly to the measure, there is a direct relationship between the measure and patient outcomes and there is a strong relationship.
** Very strong evidence to support the need for follow-up care both to improve mental health outcomes as well as to improve physical health issues.
** evidence presented, also has good "clinical" face validity re: follow-up within one month of new rx.
**agree that it's a process measure and meets moderate criteria.

** This is a process measure of adults prescribed antipsychotic medications who have a f/u appointment within 28 days. A f/u appointment allows the provider to evaluate any physical problems or complications experienced, as well as monitor effectiveness of medication.
** Yes, good evidence and practice to have timely follow up, not only to titrate medication for desired effects but also timely evaluation for side-effects that potentially impact health and compliance.
** They show evidence of low follow up rates
** Per ADA, ungraded recommendation; APA provides this a moderate clinical confidence recommendation; other guidelines also have ungraded recommendation

1b. Performance Gap

Comments:

** The measure developer did demonstrate a performance gap among the 15 states from which data was extracted. Data was not able to be extracted regarding disparities but the developer did provide literature suggesting disparities exist.
** The developer presented a 2014 cross-sectional analysis using data from 9/1/13-11/30-/14 using Medicaid Analytic eXtract (MAX) and Medicare claims. Mean score was 50.1, minimum 44.5 and maximum 60.1. Stratification demonstrated that although several of the scores hovered around the mean, there was ample room for improvement. All disparities wee accounted for.
** There is a large performance gap with only approximately 50% of patients on average receiving follow-up care.
** yes, performance gap supports variation and room for improvement
** agree w/ moderate opportunity for improvement
** In data from 15 states, less than 50% of patients had a f/u appointment within 28 days. There are racial differences in treatment of schizophrenia per data presented.
** There is what looks like a modest gap but even the highest performers are only about 60% so room for all to improve.
** Current performance showed only approx. 50% follow up across 14 states
** Performance gap demonstrated across multiple state with over 300,000 prescriptions.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Difference**s;** Comparability**;** Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable):

Link A

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*

Preliminary rating for reliability:	☒ **High**	☐ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability

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.**

Measure Number: 3313

Measure Title: Follow-Up Care for Adult Beneficiaries Who are Newly Prescribed an Antipsychotic Medication

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?
   NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

   TIPS: Consider the following: 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?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒Yes (go to Question #6)

   ☐No (please explain below then go to Question #8)

   Signal-to-noise reliability is estimated for the measure using the beta-binomial model.

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

   Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☒High (go to Question #8)

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

   The SNR analyses show that the reliability of the measure is good, and the high SNR indicates the measure performance results are highly precise with a very low degree of measurement error.

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☒Yes (go to Question #9)

   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☒Yes (go to Question #10)

   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒Yes (go to Question #2)

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

   ☒Yes (please explain below then go to Question #3)

   ☐No (go to Question #3)

   ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

   Impact of the two exclusion rules (hospitalization and death) tested across four scenarios. Analyses demonstrate the exclusions have limited impact on denominator size and measure performance.

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☐Yes (please explain below then go to Question #4)

☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

   ☐Yes (please explain below then go to Question #5)

   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

   ☐Yes (please explain below then go to Question #6)

   ☐No (go to Question #6)

   ☒Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?

   ☐Yes (please explain below then go to Question #7)

   ☒No (go to Question #7)

   Missing data analyses include all data elements required to calculate the measures: dates of service, date of birth, Medicaid eligibility, prescription fill date, National Drug Code, and type of service have negligible missingness for states in study sample.

ASSESSMENT OF MEASURE TESTING

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

   ☐Yes (go to Question #9)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☒Yes (go to Question #11)

    ☐No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☒Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

    Convergent validity was examined at state-level performance of the measure compared to several other measures of similar concepts based on TEP input.

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #14)

    ☒Moderate (go to Question #14)

    ☐Low (please explain below then go to Question #13)

    ☐Insufficient

13. Was other validity testing reported?

    ☐Yes (go to Question #14)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

    TIPS: Prior validity studies of the same data elements may be submitted

    ☒Yes (go to Question #15)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #16)

☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

Comments:

** Although doubtful to influence results, inclusion of compazine as an antipsychotic did not make sense. This medication is not used to treat psychotic illness.
** No issues with reliability. No concerns about the likelihood that this measure can be consistently applied.
** I am unclear about the numerator. Can the follow up visit be with any prescribing provider or doe sit need to be with the same provider that wrote the new prescription?
Also, there are no specifications concerning the content of the follow up visit.
**specifications are very clear. I understand they included Compazine in the list of ap meds because a phenothiazine, but this is rarely used to treat psychosis and is much more often used for nausea. For this med, can the data programmer exclude persons that had no psychiatric diagnosis?
** data elements look inclusive, no concerns
** It is not clear if a f/u visit must be face to face.
** No concerns.
** Since coming from claims should be reliable
** Good with specifications.

2a2. Reliability – Testing

Comments:

**Developer calculated state-specific measure variance as function of the measure rate at the state level as well as temporal consistency over four quarters of 2014 for entire sample by state. The signal to noise ratio was high. Reliability seems acceptable.
** No
** No concerns
** based on signal to noise ratio
** reliability results looked high to me rather than moderate, seem to explain nearly all of the state to state variability w/ little due to measurement error.
** No.
** No concerns
** Signal to noise reliability with avg reliability score of 0.98 (range: 0.88 to 0.99) - high reliability.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences

Comments:

** Face validity was good. Convergent validity was also used. This measurement made me wonder if the process being measured really had more to do with the system of care in a state rather than this specific construct--e.g. the other measures used have been in use for some time and to what extent have they fostered improvement in the process. Missing data did not seem to pose a threat to validity.
** No concerns
** I would assume that missing data problems would be addressed in the data analysis and mentioned when reporting the adherence on this measure
** no concerns, moderate makes sense given the attempt to correlate to other like measures though sounds like results were not statistically significant and correlation on graphs wasn't overly impressive.
** No concerns
** 10% missing data in NY could be due to APG rates in OP MH clinics. Might not actually pick up CPT codes. Need to check this out
** Convergent validity - found states had similar scores for other f/u related metrics (e.g. post hospitalization, antidepressant med management, start of ADHD drugs). Results: moderate for validity. Face validity looked good.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment:

Comments:

**Exclusions seem consistent with standard of care and had small influence on results.
** Denominator exclusions should be reconsidered. A hospitalization or death is exactly the outcome the measure is working to avoid.
** did not propose risk adjustment
** Exclusions are clear.
** I did not see any risk adjustment for say rural areas where there is a lack of Transportation and difficulty getting to a provider.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

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.

- All data elements are in defines fields in electronic claims.
- No fees or licensing are required.

Questions for the Committee:

- *Is the data collection strategy ready to be put into operational use?*
- *Are there any additional considerations for implementing this measure?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

Comments:

** Electronically extracted; feasible
** Claims data.-- very feasible
** CMS demonstrates feasibility using existing Medicaid claims data
** yes, feasible sources of data
** Need to clarify if only face to face visits included. How will data be gathered to assure all components of f/u visit are met, i.e., how are side effects evaluated; how is medication adherence evaluated?
** Yes, claims at state level.
** All data accessible electronically. Highly feasible.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

CMS is considering implementation options for this measure under the Innovation Accelerator Program. This measure is intended for voluntary use by states to monitor and improve the quality of care provided for Medicaid beneficiaries with physical and mental health integration needs.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

N/A

Additional Feedback (including NQF member support/non-support of the measure):

N/A

***Questions for the Committee**:*

- *Can the performance results be used to further the goal of high-quality, efficient healthcare?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results

This measure is being considered for initial endorsement. Adoption of this performance measure has the potential to improve the quality of care for Medicaid beneficiaries who are newly prescribed antipsychotic medications. On average states are providing follow-up care within four weeks for only about half of new antipsychotic prescriptions (50.2 percent). These findings suggest room for improvement in follow-up care in the states included in testing.

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation

N/A

Potential harms

N/A

Additional Feedback:

***Questions for the Committee**:*

- *How 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?*

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

** Potential use for Innovation Accelerator Program rationale explained. There is a leap being made to assume that the needed components of adequate follow-up care are being addressed if there is a visit with a prescriber.
** Unknown
** If used at the systems level it is useful, some concerns about individual provider accountability.
** easy to use
** Yes
** First submission of measure so not much feedback.
** CMS considering for duals, LTSS and substance abuse disorders.

4b1. Usability – Improvement

Comments:

** No unintended consequences evident.
** CMS is intending to use this on a voluntary basis by state as a quality improvement measure. Given potential side effects profile of antipsychotic medications, use of this measure to raise awareness about close follow up is significant.
** Benefits outweigh any potential harm
** Inclusion of compazine may inadvertently capture persons who receive this medication for only acute nausea?
** little risk of harm

** Measure has value if all components of f/u visit are met.
** I don't see any unintended consequence or harm. May put more emphasis on the importance of follow up care as defined in the measure but impact as usual depends on what is done operationally with the information since is voluntary reporting.
** This should be a useful measure
** Don't recognize any harm with measure.

Criterion 5: Related and Competing Measures

Related or competing measures
0108 : Follow-Up Care for Children Prescribed ADHD Medication (ADD)
Harmonization
Measures are completely harmonized to extent possible with same follow-up period and look-back to establish a “new prescription”.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January 10, 2018

- No comments received.
- Of the 1 NQF member who submitted a support/non-support choice:
    1. 1 supports the measure
    2. 0 do not support the measure

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

Follow_Up_New_Presc_Antipsych_Evidence_Attachment.pdf

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

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

Measure Title: Follow-Up Care for Adult Medicaid Beneficiaries Who are Newly Prescribed an Antipsychotic Medication

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

Date of Submission: 11/1/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- 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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: Follow-up care provided after a new antipsychotic prescription

☐ Appropriate use measure:

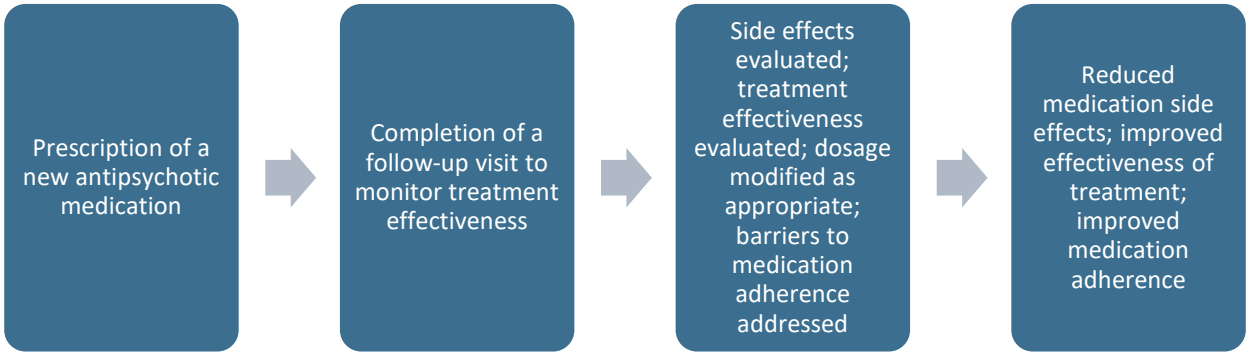
☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

Among individuals with serious mental illness, physical health problems such as cardiovascular disease, metabolic disorders, and infectious disease are more prevalent compared to the general population. Antipsychotic medications can exacerbate existing physical problems as well as increase a patient's risk for developing new health concerns such as metabolic complications. Timely follow-up with a provider following the prescription of antipsychotic medications is an essential first step to ensure that physical impacts of antipsychotic medications are identified and addressed early. Early follow up is also critical to monitor for treatment effectiveness and modify dosage as necessary, as well as to identify and address any barriers to treatment adherence. By proactively following up with patients who are prescribed antipsychotic medications, providers can identify problems early in the course of treatment and minimize potential

harms associated with use of those medications. Regardless of the care setting in which a patient is being treated, comprehensive assessment of both physical and mental health factors is an essential aspect of treatment with antipsychotic medications.



1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

N/A; not derived from patient report

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

N/A; not an outcome measure.

1a.3. SYSTEMATIC REVIEW(S) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Taken together, the four guidelines cited below provide recommendations that emphasize the importance of ongoing, comprehensive assessment of both physical and mental health factors for patients taking psychotropic medications, with Guidelines 1 and 2 focusing specifically on antipsychotic medications. Each guideline provides an overview of key factors that need to be assessed as part of the course of treatment. A comprehensive assessment of physical and mental

health factors is an essential part of treatment for patients who are prescribed psychotropic medications, regardless of the care setting in which their treatment is provided. Cited below are the specific guideline recommendations around comprehensive assessment for patients taking antipsychotic medications and/or psychotropic medications.

Guideline 1.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**American Diabetes Association (ADA), American Psychiatric Association (APA), American Association of Clinical Endocrinologists (AACE), North American Association for the Study of Obesity (NAASO). "Consensus development conference on antipsychotic drugs and obesity and diabetes."** ***Diabetes Care.*** **2004;27(2): 596-601. Available at http://care.diabetesjournals.org/content/27/2/596**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**Follow-up monitoring:** "The patient's weight should be reassessed at 4, 8, and 12 weeks after initiating or changing second generation antipsychotic (SGA) therapy and quarterly thereafter at the time of routine visits." "Fasting plasma glucose, lipid levels, and blood pressure should also be assessed 3 months after initiation of antipsychotic medications. Thereafter, blood pressure and plasma glucose values should be obtained annually or more frequently in those who have a higher baseline risk for the development of diabetes or hypertension. In those with a normal lipid profile, repeat testing should be performed at 5-year intervals or more frequently if clinically indicated" "Blood pressure, lipid, and glycemic goals of therapy for people with diabetes apply equally to those who also have psychiatric disorders. However, all goals need to be individualized. The benefits and risks of different therapeutic agents used in the treatment of diabetes and its comorbidities should be considered in the context of the patient's psychiatric condition and treatment. In summary, the panel recommends the following: • Consideration of metabolic risks when starting SGAs • Patient, family, and care giver education • Baseline screening • Regular monitoring • Referral to specialized services, when appropriate"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline did not assign a grade to the quality of the quoted evidence.
Provide all other grades and definitions from the evidence grading system	None.
Grade assigned to the **recommendation** with definition of the grade	The guideline did not provide a grade for the cited recommendations.
Provide all other grades and definitions from the recommendation grading system	None.

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	This guideline was developed as a result of a consensus development conference of key experts and stakeholders. The key goal of the conference was to establish consensus on the following questions: 1. What is the current use of antipsychotic drugs? 2. What is the prevalence of obesity, pre-diabetes, and type 2 diabetes in the populations in which second-generation antipsychotics (SGAs) are used? 3. What is the relationship between the use of SGAs and the incidence of obesity or diabetes? 4. Given the above risks, how should patients be monitored for the development of significant weight gain, dyslipidemia, and diabetes, and how should they be treated if diabetes develops? 5. What research is needed to better understand the relationship between these drugs and significant weight gain, dyslipidemia, and diabetes? While this is a consensus-based guideline, the authors cited 4 clinical practice guidelines, 5 systematic evidence reviews, 11 retrospective analyses, 3 randomized trials, and 1 cross-sectional analysis in support of the document. Evidence cited in support of the consensus document ranges from 1997-2003.
Estimates of benefit and consistency across studies	The guideline was developed using a consensus-based approach involving a group of stakeholders and experts in the field. While the consensus-based approach was supplemented with a review of the evidence as part of the development of each guideline, details of the consistency of the reviewed evidence were not provided. This guideline does not provide a quantitative estimate of benefit for follow-up care for patients prescribed antipsychotic medications. However, there is consensus among the guidelines that close follow-up monitoring is an essential standard of care for patients prescribed antipsychotic medications to ensure effectiveness of treatment and to mitigate any adverse consequences or reactions to the drugs.
What harms were identified?	While the guideline does not give a formal description or estimate of harms, we anticipate the expected benefits of follow-up care to far outweigh any potential harms because ongoing monitoring and follow-up is a basic standard of care for patients taking antipsychotic medications.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We did not identify any new studies since the clinical guideline was published that change the conclusion in the recommendation.

Guideline 2.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Lehman AF, Lieberman JA, Dixon LB, et al; American Psychiatric Association; Work Group on Schizophrenia. "Practice guideline for the treatment of patients with schizophrenia, second edition."** ***Am J Psychiatry.*** **2004;161(2 Suppl):1-56. Available at** **https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	"The recommended dose is that which is both effective and not likely to cause side effects that are subjectively difficult to tolerate, since the experience of unpleasant side effects may affect long-term adherence [I]. The dose may be titrated as quickly as tolerated to the target therapeutic dose of the antipsychotic medication, and unless there is evidence that the patient is having uncomfortable side effects, monitoring of the patient's clinical status for 2–4 weeks is warranted to evaluate the patient's response to the treatment [II].**"**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline did not assign a grade to the quality of the quoted evidence.
Provide all other grades and definitions from the evidence grading system	None.
Grade assigned to the **recommendation** with definition of the grade	The guideline assigned a grade II to its follow-up recommendation.
Provide all other grades and definitions from the recommendation grading system	The APA grading scale is defined as follows: [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence. [III] May be recommended on the basis of individual circumstances.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The evidence reviewed to support this guideline included clinical trials and meta-analyses related to schizophrenia and schizoaffective disorder to reflect a synthesis of the current literature and clinical practice on the treatment of patients with schizophrenia. Evidence cited in support of this guideline includes 181 double-blind randomized clinical trials, 116 randomized clinical trials, 152 clinical trials, 133 cohort or longitudinal studies, 122 case-control studies, 71 reviews with secondary data analysis, 167 literature reviews, and 497 other types of studies. Evidence reviewed in the development of this guideline spanned from 1994-2002.

Estimates of benefit and consistency across studies	The APA guideline was developed based on a comprehensive literature review. However, details of the consistency of the reviewed evidence were not provided. The guideline does not provide a quantitative estimate of benefit for follow-up care for patients prescribed antipsychotic medications. However, there is consensus among the guidelines that close follow-up monitoring is an essential standard of care for patients prescribed antipsychotic medications to ensure effectiveness of treatment and to mitigate any adverse consequences or reactions to the drugs.
What harms were identified?	While the guideline does not give a formal description or estimate of harms, we anticipate the expected benefits of follow-up care to far outweigh any potential harms because ongoing monitoring and follow-up is a basic standard of care for patients taking antipsychotic medications.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We did not identify any new studies since the clinical guideline was published that change the conclusion in the recommendation.

Guideline 3.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**University of South Florida College of Behavioral and Community Sciences. "2015 Florida Best Practice Psychotherapeutic Medication Guidelines for Adults." The University of South Florida, Florida Medicaid Drug Therapy Management Program sponsored by the Florida Agency for Health Care Administration, December 2015. Available at: http://www.medicaidmentalhealth.org/_assets/file/Guidelines/Web_2015-Psychotherapeutic%20Medication%20Guidelines%20for%20Adults_Final_Approved1.pdf**

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**Principles of Practice** **Comprehensive Assessment** • Careful, differential diagnostic evaluation • Risk for suicide and violence • Co-occurring mental and medical disorders • Substance abuse disorders, including tobacco use • Potential bipolar disorder must be assessed in patients presenting with depression • Serious mental health conditions are chronic in nature; therefore, a long-term management plan is essential o Use measurement-based care to measure symptoms, side effects, and adherence o Select maintenance medications that have a low relative risk of weight gain and metabolic syndrome o Monitoring of physical health parameters and medication side effects (See Program publication A Summary for Monitoring Physical Health and Side-Effects of Psychiatric Medications in the Severely Mentally Ill Population available at www.medicaidmentalhealth.org) o Integrate care of psychiatrists and primary care providers o Incorporate collaborative/shared treatment decision-making with patients and family/caregivers o Perform a psychosocial assessment o Assess social support system (housing, family, other caregivers) o Evaluate threats to continuity of care (access to medication, adherence, etc.) o Give patients tools/support for recovery and self-management **Adjunctive Psychosocial Treatments (As Indicated)** • Individual and family psychoeducation • Cognitive-behavioral therapy (CBT) • Interpersonal psychotherapy (IPT) • Interpersonal and social rhythm therapy (IPSRT) • Family-focused therapy • Group psychoeducation (especially for bipolar disorder) • Social skills training (especially in schizophrenia)

	• Cognitive remediation/rehabilitation (to improve attention, memory, and/or executive function) **Note on pharmacogenomic testing - Limited data exists examining whether patient care that integrates pharmacogenomic test information results in better or safer treatment.* **Measurement-Based Care** Questionnaires and rating scales are useful tools for diagnostic assessment and evaluation of treatment outcomes, and such instruments can be helpful in providing supplemental information to clinical judgment. The integration of measurement scales into routine clinical practice is suggested for each of the conditions covered in this document. Clinicians should use rating scales to assess symptom severity during the initial evaluation/treatment, when medication changes are implemented, and/or when the patient reports a change in symptoms. • Treatment targets need to be precisely defined. • Effectiveness and safety/tolerability of the medication treatment must be systematically assessed by methodical use of appropriate rating scales and side-effect assessment protocols. • Internet links to the following scales are available on the program website -www.medicaidmentalhealth.org o Beck Depression Inventory (BDI) o Brief Psychiatric Rating Scale (BPRS) o Clinical Global Impression (CGI) Scale o Clinician-Rated Dimensions of Psychosis Symptom Severity (CRDPSS) o Hamilton Rating Scale for Depression (HAM-D) o Montgomery-Asberg Depression Rating Scale (MADRS) o Patient Health Questionnaire (PHQ-9) o Positive and Negative Syndrome Scale (PANSS) o Quick Inventory of Depression Symptomatology (QIDS) o Young Mania Rating Scale (YMRS)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline did not assign a grade to the quality of the quoted evidence.
Provide all other grades and definitions from the evidence grading system	None.
Grade assigned to the **recommendation** with definition of the grade	The guideline did not provide a grade for the cited recommendations.

Provide all other grades and definitions from the recommendation grading system	None.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The guideline was developed using a consensus-based approach. While evidence was reviewed as part of the development process for these guidelines, the details of the evidence review were not provided.
Estimates of benefit and consistency across studies	The guideline was developed using a consensus-based approach involving a group of stakeholders and experts in the field. While the consensus-based approach was supplemented with a review of the evidence as part of the development of each guideline, details of the consistency of the reviewed evidence were not provided. The guideline does not provide a quantitative estimate of benefit for follow-up care for patients prescribed antipsychotic medications. However, there is consensus among the guidelines that close follow-up monitoring is an essential standard of care for patients prescribed antipsychotic medications to ensure effectiveness of treatment and to mitigate any adverse consequences or reactions to the drugs.
What harms were identified?	While the guideline does not give a formal description or estimate of harms, we anticipate the expected benefits of follow-up care to far outweigh any potential harms because ongoing monitoring and follow-up is a basic standard of care for patients taking antipsychotic medications.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We did not identify any new studies since the clinical guideline was published that change the conclusion in the recommendation.

Table 4.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**University of South Florida College of Behavioral and Community Sciences. "A Summary for Monitoring Physical Health and Side-Effects of Psychiatric Medications in the Severely Mentally Ill Population." The University of South Florida, Florida Medicaid Drug Therapy Management Program for Behavioral Health sponsored by the Florida Agency for Health Care Administration, March 2014. Available at: http://medicaidmentalhealth.org/_assets/file/Summaries/2014_Monitoring%20Physical%20Health%20and%20Side-Effects%20of%20Psychiatric%20Medicati....pdf**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	See Table 12 below.

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline did not assign a grade to the quality of the quoted evidence.
Provide all other grades and definitions from the evidence grading system	None.
Grade assigned to the **recommendation** with definition of the grade	The guideline did not provide a grade for the cited recommendations.
Provide all other grades and definitions from the recommendation grading system	None.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The guideline was developed using a consensus-based approach. While evidence was reviewed as part of the development process for these guidelines, the details of the evidence review were not provided.
Estimates of benefit and consistency across studies	The guideline was developed using a consensus-based approach involving a group of stakeholders and experts in the field. While the consensus-based approach was supplemented with a review of the evidence as part of the development of each guideline, the consistency of the reviewed evidence were not provided. None of the cited guidelines provide a quantitative estimate of benefit for follow-up care for patients prescribed antipsychotic medications. However, there is consensus among the guidelines that close follow-up monitoring is an essential standard of care for patients prescribed antipsychotic medications to ensure effectiveness of treatment and to mitigate any adverse consequences or reactions to the drugs.
What harms were identified?	While the guideline does not give a formal description or estimate of harms, we anticipate the expected benefits of follow-up care to far outweigh any potential harms because ongoing monitoring and follow-up is a basic standard of care for patients taking antipsychotic medications.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We did not identify any new studies since the clinical guideline was published that change the conclusion in the recommendation.

General Recommendations: Monitoring Physical Health in Patients with Severe Mental Illness

Table 12: Monitoring Patients with Severe Mental Illnesses: Recommended Frequency of Assessment[1,2]

Assessment	Baseline	During titration/ At target dose	Each Visit	At 6 weeks	At 3 months	Every 3 months	At 12 months	Annually after first 12 months
Personal and family history	✓	–	–	–	–	–	✓	✓
Lifestyle behaviors (smoking, exercise, dietary habits)	✓	–	✓	✓	✓	–	✓	✓
Weight	✓	–	✓	✓	✓	–	✓	✓
Waist circumference*	✓	–	–	✓	✓	–	✓	✓
BP and pulse	✓ (during titration with clozapine and quetiapine)	✓	✓	✓	✓	–	✓	✓
Sedation/somnolence	✓	–	✓	✓	✓	–	✓	✓
Sexual/reproductive dysfunction	✓	✓	–	✓	✓	✓	–	–
Prolactin	✓a	–	–	–	✓b	–	✓b	✓b
Fasting blood glucose	✓	–	–	✓	✓	–	✓	✓
Fasting lipid profile	✓	–	–	–	✓	–	✓	✓
Parkinsonism (SAS or ESRS), Akathisia (AIMS or ESRS)†	✓	✓	–	–	✓	–	✓	✓
Electrolytes, full blood count, renal function	✓	–	–	–	–	–	✓(more frequent blood counts if on clozapine)	✓(more frequent blood counts if on clozapine)
Ftardive dyskinesia	✓	–	–	–	–	–	✓	✓
Liver function tests	✓	–	–	–	–	–	✓	✓
Dental Health ECG‡ parameters	✓‡	–	–	–	–	–	–	–

*Studies have shown that waist circumference is a better predictor of cardiovascular risk compared to Body Mass Index (BMI)

[a]recommended to obtain baseline values; if too expensive, obtain only in cases where sexual or reproductive system abnormalities are reported

[b]obtain in cases where sexual dysfunction coincides with antipsychotic treatment or dose change

‡ECG = electrocardiogram; perform EKG at baseline then only if symptomatic

[1]Adapted from Hert, et al, 2011. "Physical Illness in patients with severe mental disorders, II, Barriers to care, monitoring, and treatment guidelines, plus recommendations at the system and individual level. World Psychiatry. 10: 138-151.

[2]Adapted from Florida Medicaid Drug Therapy Management Program for Behavioral Health: Florida Best Practice Medication Child/ Adolescent Guidelines

†Abbreviations: SAS = Simpson-Angus Scale; ESRS = Extrapyramidal Symptom Rating Scale; AIMS = Abnormal Involuntary Movement Scale.

_____

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

In addition to the clinical guidelines reviewed above, the project team reviewed literature and consulted with stakeholders and clinical experts in development of this measure (see below). To define the measure specifications, the project team convened and consulted with a clinical advisory workgroup, including the follow-up period, target medications, and types of follow-up visits. The workgroup noted that timely follow-up is a minimal clinical standard of care for patients with mental illness who are prescribed antipsychotics and other psychotropic medications and is a critical component of disease management.

The literature also supports the underlying concept that follow-up visit with a provider is essential to monitor treatment effectiveness, evaluate health concerns, and adjust treatment as needed to minimize potential harms associated with the use of psychotropic medications. One 2014 cross-sectional analysis of nationally-representative data estimates that 35 percent to 50 percent of mental health care episodes consist of psychotropic drug fills without an outpatient visit to monitor treatment and up to 35 percent of episodes consisted of only a single visit (Le Cook et al. 2014).

Despite the importance of follow-up for patients taking antipsychotics, there is evidence that these patients are not receiving adequate follow-up care. For example, there is a growing body of evidence that shows persistent gaps in monitoring for metabolic effects of antipsychotic medications despite available guidelines and recommendations. While follow-up care should encompass more than just metabolic monitoring, metabolic testing rates can be useful to gain a general idea of the adequacy of follow-up care. In a 2016 analysis of data from the Missouri Medicaid program, Morrato and colleagues reported annual testing rates of 79.6 percent for glucose and 41.2 percent for lipids among beneficiaries taking antipsychotics (Morrato et al. 2016). This shows improvement over an earlier 2010 analysis data from three state Medicaid programs, which found testing rates as low as 27 percent for glucose testing and 10 percent for lipid testing (Morrato et al. 2010). This improvement is consistent with a 2011 analysis of Kansas Medicaid data that found improvement in annual testing between 2002 and 2007 from 23 percent to 75.3 percent for glucose monitoring and from 10.1 percent to 52.5 percent for lipid monitoring (Moeller, Rigler, Mayorga, Nazir, & Shireman 2011).

While progress on testing at a state level is encouraging, there is still considerable room for improvement at a local level. In a 2011 analysis of Medicaid data, rates of metabolic testing were found to vary significantly based on geographic location and patient characteristics such as age and comorbidity (Morrato et al. 2011). Inadequate follow-up care is often reflected by poor treatment adherence. A 2015 study found irregular attendance at follow-up appointments to be significantly associated with medication nonadherence (OR: 5.7; 95 percent confidence interval 2.92-11.31) among patients with psychiatric illness (Mert et al. 2015). A 2013 study found an antipsychotic non-adherence rate of nearly 38 percent among Medicaid patients with schizophrenia, with new prescription of antipsychotics and baseline non-adherence increasing the likelihood of non-adherence twelvefold (Lang et al. 2013). Appropriate follow-up care is essential for patients taking antipsychotic medications to receive the full benefit of treatment and to minimize potential harms associated with use of antipsychotics.

The importance of ongoing follow-up for patients on psychotropic medications is also emphasized in recent government efforts to promote best prescribing practices for psychotropic medications (MACPAC 2015). In a 2015 study, Mert and

colleagues identified irregular follow-up as an important risk factor for medication non-adherence among patients with mental illness (Mert et al. 2015).

1a.4.2 What process was used to identify the evidence?

The project team conducted an environmental scan, which included a targeted literature review, an evaluation of existing performance measures related to physical and mental health care integration to identify critical measurement gaps, and interviews with key stakeholders and subject matter experts. Stakeholders interviewed by the project team emphasized the importance of ongoing follow-up after the prescription of psychotropic medications to evaluate treatment effectiveness and modify the treatment regimen as appropriate. Timely follow-up is also essential to address medication side effects and potential barriers to treatment adherence. As noted above, the project team consulted with a clinical advisory work group to identify the appropriate follow-up period, types of follow-up visits, and types of medications for inclusion in the measure.

1a.4.3. Provide the citation(s) for the evidence.

De Hert M, Correll CU, Bobes J, et al. Physical illness in patients with severe mental disorders. I. Prevalence, impact of medications and disparities in health care. *World Psychiatry.*2011;10:52-77.

Lang K, Federico V, Muser E, Menzin J, Menzin J. Rates and predictors of antipsychotic non-adherence and hospitalization in Medicaid and commercially-insured patients with schizophrenia. J Med Econ. 2013;16(8):997-1006. doi: 10.3111/13696998.2013.816310.

Le Cook B, Zuvekas SH, Carson N, et al. Assessing racial/ethnic disparities in treatment across episodes of mental health care. *Health Serv Res.*2014;49(1):206-29. doi: 10.1111/1475-6773.12095.

Medicaid and CHIP Payment and Access Commission (MACPAC). Report to Congress on Medicaid and CHIP. June 2015. available at: https://www.macpac.gov/wp-content/uploads/2015/06/June-2015-Report-to-Congress-on-Medicaid-and-CHIP.pdf. Accessed February 4, 2016.

Mert DG, Turgut NH, Lelleci M, Semiz M. Perspectives on reasons of medication nonadherence in psychiatric patients. Patient Prefer Adherence. 2015;9:87-93. doi: 10.2147/PPA.S75013.

Moeller KE, Rigler SK, Mayorga A, Nazir N, and Shireman TI. Quality of monitoring for metabolic effects associated with second generation antipsychotics in patients with schizophrenia on public insurance. *Schizophr Res.* 2011;126(1-3):117-23. doi: 10.1016/j.schres.2010.11.015

Morrato EH, Campagna EJ, Brewer SE, et al. Metabolic testing for adults in a state Medicaid program receiving antipsychotics: remaining barriers to achieving population health prevention goals. *JAMA Psychiatry.*2016;73(7):721-30. doi:10.1001/jamapsychiatry.2016.0538

Morrato EH, Druss B, Hartung DM, et al. Metabolic testing rates in 3 state Medicaid programs after FDA warnings and ADA/APA recommendations for second-generation antipsychotic drugs. *Arch Gen Psychiatry.* 2010;67(1):17-24. doi:10.1001/archgenpsychiatry.2009.179

Morrato EH, Druss BG, Hartung DM, et al. Small area variation and geographic and patient-specific determinants of metabolic testing in antipsychotic users. *Pharmacoepidemiol Drug Saf.* 2011;20(1):66-75. doi:10.1002/pds.2062.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Among individuals with serious mental illness, physical health problems such as cardiovascular disease, metabolic disorders, and infectious disease are more prevalent compared to the general population. Antipsychotic medications can exacerbate existing physical problems as well as increase a patient's risk for developing new health concerns such as metabolic complications. Timely follow-up with a provider following the prescription of antipsychotic medications is an essential first step to ensure that physical impacts of antipsychotic medications are identified and addressed early. Early follow up is also critical to monitor for treatment effectiveness and modify dosage as necessary, as well as to identify and address any barriers to treatment adherence. By proactively following up with patients who are prescribed antipsychotic medications, providers can identify problems early in the course of treatment and minimize potential harms associated with use of those medications. Regardless of the care setting in which a patient is being treated, comprehensive assessment of both physical and mental health factors is an essential aspect of treatment with antipsychotic medications.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The measure was tested using Medicaid Analytic eXtract (MAX) and Medicare claims data from September 1, 2013 to November 30, 2014. The measurement period was eleven months (January 1 through November 30, 2014) to allow for the four week follow-up period. In addition, the data allowed for a four month look-back into the prior year to ensure continuous enrollment (September 1 through December 2013).

The measure was tested for 15 states: Arkansas, Connecticut, Georgia, Iowa, Michigan, Mississippi, Missouri, New Jersey, New York, Pennsylvania, South Dakota, Tennessee, Vermont, West Virginia, and Wyoming.

For Medicaid-only and dual-eligible beneficiaries over age 18 years in our sample states, there were 3,768,880 antipsychotic prescriptions filled in calendar year 2014; 332,736 of these prescriptions were filled during our measure period (January 1, 2014 through November 30, 2014) and met our definition of a "new antipsychotic prescription." Of these, 267,831 prescriptions were for beneficiaries who met the continuous enrollment requirements and the inpatient hospitalization and death exclusions. A total of 130,785 (49 percent) of these prescriptions were associated with follow-up care within four weeks. Across all states, measure performance (that is, the proportion of prescriptions filled by individuals who received a follow-up visit within four weeks of their fill date) was 48.8 percent (130,785 follow-ups to 267,831 new prescriptions).

All States:

Mean: 50.2%

Std. Deviation: 5.3%

Min: 44.5%

25th Percentile: 46%

Median: 47.9%

75th Percentile: 54.3%

Max: 60.1%

Interquartile Range: 8.2%

Below we present the number of patients who met the measure's criteria for denominator and numerator by state, along with each state's measure performance rate and confidence intervals relative to overall performance rate.

Arkansas:

-Denominator: 5,751

-Numerator: 2,558

-Performance rate: 44.48%

-CI: 43.19%, 45.76%

Connecticut:

-Denominator: 13,150

-Numerator: 6,234

-Performance rate: 47.41%

-CI: 46.55%, 48.26%

Georgia:

-Denominator: 18,705

-Numerator: 8,473

-Performance rate: 45.30%

-CI: 44.58%, 46.01%

Iowa:

-Denominator: 8,705

-Numerator: 4,986

-Performance rate: 57.28%

-CI: 56.24%, 58.32%

Michigan:

-Denominator: 29,545

-Numerator: 17,177

-Performance rate: 58.14%

-CI: 57.58%, 58.7%

Mississippi:

-Denominator: 18,025

-Numerator: 8,883

-Performance rate: 49.28 %

-CI: 44.37%, 46.47%

Missouri:

-Denominator: 8,640

-Numerator: 3,924

-Performance rate: 45.42%

-CI: 48.55%, 50.01%

New Jersey:

-Denominator: 17,937

-Numerator: 8,420

-Performance rate: 46.94%

-CI: 46.21%, 47.67%

New York:

-Denominator: 75,644

-Numerator: 36,219

-Performance rate: 47.88%

-CI: 47.52%, 48.24%

Pennsylvania:

-Denominator: 41,979

-Numerator: 19,597

-Performance rate: 46.68%

-CI: 46.21%, 47.16%

South Dakota:

-Denominator: 1,141

-Numerator: 582

-Performance rate: 51.01%

-CI: 48.11%, 53.91%

Tennessee:

-Denominator: 19,269

-Numerator: 8,666

-Performance rate: 44.97%

-CI: 44.27%, 45.68%

Vermont:

-Denominator: 2,525

-Numerator: 1,401

-Performance rate: 55.49%

-CI: 53.55%, 57.42%

West Virginia:

-Denominator: 6,083

-Numerator: 3,225

-Performance rate: 53.02%

-CI: 51.76%, 54.27%

Wyoming:

-Denominator: 732

-Numerator: 440

-Performance rate: 60.11%

-CI: 56.56%, 63.66%

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.

Not applicable. Data have been included in Section 1b.2.

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out",*

disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.

During testing, the measure performance was stratified for disparities by age, gender, race, disability status, and beneficiary category (Medicaid-only or dually eligible for Medicaid and Medicare).

Measure Performance by Age:

18-34

-Denominator: 65,162

-Numerator: 32,746

-Performance Rate: 50.25%

35-44

-Denominator: 45,932

-Numerator: 24,738

-Performance Rate: 53.86%

45-54

-Denominator: 59,976

-Numerator: 32,348

-Performance Rate: 53.93%

55-64

-Denominator: 44,690

-Numerator: 22,945

-Performance Rate: 51.34%

65-74

-Denominator: 23,067

-Numerator: 9,438

-Performance Rate: 40.92%

75+

-Denominator: 29,004

-Numerator: 8,570

-Performance Rate: 29.55%

Measure Performance by Race (All Ages):

White

-Denominator: 152,748

-Numerator: 76,836

-Performance Rate: 50.30%

Black

-Denominator: 70,886

-Numerator: 32,135

-Performance Rate: 45.33%

Hispanic/Latino:

-Denominator: 27,838

-Numerator: 13,161

-Performance Rate: 47.28%

Other or unknown race/ethnicity

-Denominator: 16,359

-Numerator: 8,653

-Performance Rate: 52.89%

Measure Performance by Medicaid Beneficiary Category (All Ages):

Medicaid-only

-Denominator: 132,835

-Numerator: 75,667

-Performance Rate: 56.96%

Dually eligible for Medicaid and Medicare

-Denominator: 134,996

-Numerator: 55,118

-Performance Rate: 40.83%

Source: Mathematica analysis of 2013–2014 MAX PS, RX, OT, and IP files and Medicare Part B, Part D, OPD, and MedPAR files.

Notes: Data do not include patients with missing or unknown characteristics data.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

Please see data provided in 1b.4. Additionally, we provide a brief summary of the literature below.

The literature demonstrates racial and ethnic differences in the way patients seek treatment for mental illness as well as the way mental illness is managed. A 2014 cross-sectional analysis of nationally-representative data found that black and Latino patients were less likely than white patients to initiate treatment and receive adequate treatment for mental illness. Black patients were also more likely to have an episode of care that included a psychiatric emergency department or inpatient visit. However, white patients were more likely than Latino or black patients to experience an episode of care with that included continuous psychotropic drug fills without an outpatient visit to monitor treatment (Le Cook et al. 2014). A 2014 analysis of Medicaid claims data from California, Florida, New York, and North Carolina found that that black and Latino beneficiaries experienced poorer quality schizophrenia care than white beneficiaries, as measured by a composite measure of quality derived from 14 evidence-based quality indicators. In particular, black and Latino patients had lower scores on metrics such as antipsychotic adherence, psychosocial visits, routine psychotherapy, routine psychiatric care, and follow-up after discharge, with Latino patients generally experiencing better care than blacks but worse care than whites (Horvitz-Lennon et al. 2014).

Horvitz-Lennon M, Volya R, Donohue JM, Lave JR, Stein BD, Normand SLT. Disparities in quality of care among publicly insured adults with schizophrenia in four large U.S. states, 2002-2008. Health Serv Res. 2014;49(4):1121-44. doi: 10.1111/1475-6773.12162.

Le Cook B, Zuvekas SH, Carson N, et al. Assessing racial/ethnic disparities in treatment across episodes of mental health care. Health Serv Res.2014;49(1):206-29. doi: 10.1111/1475-6773.12095.

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 sub criteria 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):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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.)*

The measure does not yet have published specifications; therefore no link exists.

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:** Follow_Up_New_Presc_Antipsych_Codes.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Not applicable.

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) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Antipsychotic prescriptions from the denominator prescribed to a beneficiary who completed a follow-up visit with a provider with prescribing authority within four weeks of prescription of an antipsychotic medication.

S.5. 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, time period for data collection, 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 (S.14).

The proposed numerator uses a four-week follow-up period based on clinical guidelines for appropriate follow-up after prescription of new antipsychotic medications. The optimal follow-up period was determined through testing and consultation with the Clinical Advisory Work group. The day after the prescription is counted as day 1 of the follow-up period. The date of the follow-up visit with a provider is determined by using the service date on the medical claim.

See attached Excel file for CPT and HCPCS codes that qualify for the numerator.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

New antipsychotic prescriptions for Medicaid beneficiaries age 18 years and older.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Target population meets the following conditions:

1. Medicaid beneficiary age 18 years and older (including dual-eligible and Medicaid-only enrollees)
2. Newly prescribed an antipsychotic medication
3. Enrolled in Medicaid during the four months prior to and the four weeks following a new prescription of an antipsychotic medication

Beneficiaries with "newly filled prescription" are those who have had no antipsychotic medications dispensed for either new or refill prescriptions during a period of 120 days (four months) prior to the prescription fill date.

The measure focuses on new prescriptions of antipsychotic medications.

We used National Drug Codes to identify the following antipsychotic medications for this measure:

- aripiprazole (Abilify)
- asenapine maleate (Saphris)
- chlorpromazine hydrochloride
- clozapine (Clozaril, FazaClo, Versacloz)
- Compazine
- droperidol (Inapsine)
- fluoxetine hydrochloride-olanzapine (Symbyax)
- fluoxetine-olanzapine
- fluphenazine
- haloperidol (Haldol)
- iloperidone (Fanapt)
- loxapine succinate (Loxitane)
- lurasidone hydrochloride (Latuda)
- molindone hydrochloride (Moban)
- olanzapine (Zyprexa)
- paliperidone (Invega)
- Permitil
- perphenazine

- pimozide (Orap)
- prochlorperazine maleate
- quetiapine fumarate (Seroquel)
- risperidone (Risperdal)
- thioridazine hydrochloride
- thiothixene (Navane)
- trifluoperazine hydrochloride
- trilafon
- ziprasidone (Geodon)

See attached Excel file for NDCs that qualify for the denominator.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

• Medicaid beneficiaries with an acute inpatient admission during the four-week follow-up period after prescription of an antipsychotic medication

• Patients who expired within four weeks of new prescription date.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.)*

Acute inpatient admission during the four-week follow-up period: Beneficiaries with an inpatient admission during the four week follow-up period are excluded from the measure.

Death: Patients with a date of death during the four-week follow-up period are excluded from the measure.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.)*

Not applicable; this measure is not stratified.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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 = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

To calculate the denominator:

Eligible Population:

1. Identify Medicaid beneficiaries (both dual-eligible and Medicaid-only enrollees) age 18 years and older.

2. From this group, identify those who were newly prescribed one or more antipsychotic medications.

Exclusions:

From the population identified in step 2

3. Remove any beneficiaries who were not continuously enrolled for at least four months before or four weeks following the new prescription.

4. Remove any beneficiaries who had an acute inpatient admission during the four weeks following the new prescription.

5. Remove any beneficiaries who died during the four weeks following the new prescription.

Numerator

From the beneficiaries within the denominator (after denominator exclusions have been applied)

6. Identify the number of beneficiaries who had a qualifying outpatient encounter within four weeks of the prescription date of the antipsychotic medication.

To calculate the measure score:

7. Divide the total number of beneficiaries in the numerator by the total number of beneficiaries in the denominator, after denominator exclusions have been applied.

8. Multiply this number by 100 to determine the performance rate.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Not applicable; this measure does not use a sample.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

Not applicable; this measure does not use a survey.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Medicaid and Medicare administrative claims or encounter data and pharmacy claims. Data sources include:

• State Medicaid Management Information System (MMIS), MSIS, or T-MSIS or Medicaid Analytic eXtract (MAX) file: MAX PS, MAX RX, MAX IP, MAX OT

• Additional Data Sources for dual-eligible beneficiaries: Medicare Parts A, B, and D data

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Population : Regional and State

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Outpatient Services

If other:

S.22. 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.)*

Not applicable.

2. Validity – See attached Measure Testing Submission Form

Follow_Up_New_Presc_Antipsych_Testing_Attachment.pdf

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:**
Measure Title: Follow-Up Care for Adult Medicaid Beneficiaries Who are Newly Prescribed an Antipsychotic Medication
Date of Submission: 11/1/2017

Type of Measure:

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

Instructions

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

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

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

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

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

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

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

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

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*).

For both Medicaid and dual-eligible beneficiaries, we analyzed data from the Medicaid Analytic eXtract (MAX) files, which include five types of files for each state and year as follows:

- Person Summary (PS)—Person-level file including Medicaid eligibility and demographic information
- Inpatient (IP) files—Claim-level file including information on inpatient hospital stays
- Long-Term Care (LT)—Claim-level file including information on long-term care institutional stays (nursing facilities, intermediate care facilities for individuals with intellectual disabilities, psychiatric hospitals, etc.)
- Other Therapy (OT)—Claim-level file including information on use of "other" services, such as outpatient care and home- and community-based service use
- Drug (RX) files—Claim-level information on drugs and other services provided by a pharmacy

For the beneficiaries dually eligible for both Medicare and Medicaid, we also used Medicare enrollment data from the Medicare Beneficiary Summary File (MBSF) and from the following sources of claims data in CMS's Data Extract System (DESY):

- Part D Prescription Drug claims from CMS's Prescription Drug Events file
- Carrier or Physician/Supplier Part B claims from the National Claims History (NCH) file for final action, fee-for-service (FFS) claims submitted on a CMS-1500 claim form from non-institutional providers such as physicians, physician assistants, clinical social workers, and nurse practitioners
- Outpatient Department (OPD) claims from the NCH for final action, FFS claims data submitted by institutional outpatient providers such as hospital outpatient departments, rural health clinics, renal dialysis facilities, outpatient rehabilitation facilities, comprehensive outpatient facilities, and community mental health centers
- Medicare Provider Analysis Review (MedPAR) files for final action acute hospital and skilled nursing facility (SNF) admissions

1.3. What are the dates of the data used in testing? We tested the measure (referred to here as "PMH-1") using Medicaid and Medicare data from calendar year 2014 (the measurement year). The measure requires a 16-week look-

back period to establish that a beneficiary did not have any other antipsychotic medications prescribed; thus, beneficiaries are required to be enrolled in Medicaid for the 16 weeks before the prescription was filled. To account for the look-back period and establish continuous enrollment during this period, we also used data from September 1, 2013, through December 31, 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.20***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: state	☒ other: state

We calculated state-specific measure variance ("noise") as a function of the measure rate at the state level. We assessed the temporal consistency (also referred to as temporal stability) of the PMH-1 measure by examining the strength of association between measure results in four quarters of the 2014 measurement year for the entire sample by state.

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

15 states: Arkansas, Connecticut, Georgia, Iowa, Michigan, Mississippi, Missouri, New Jersey, New York, Pennsylvania, South Dakota, Tennessee, Vermont, West Virginia, and Wyoming.

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*)

Of the Medicaid-only and dual-eligible beneficiaries over age 18 years in our sample states, 550,842 beneficiaries had at least one antipsychotic prescription in the measure year. Because the measure looks at whether follow-up occurred after a new antipsychotic prescription (that is, the measure is at the prescription rather than the beneficiary level), the remaining description focuses on the number of prescriptions rather than the number of beneficiaries. There were 3,768,880 antipsychotic prescriptions filled in calendar year 2014 by the Medicaid-only and dual-eligible beneficiaries in our sample; 332,736 of these prescriptions were filled during our measure period (January 1, 2014 through November 30, 2014) and met our definition of a "new antipsychotic prescription." Of these, 267,831 prescriptions met the continuous enrollment requirements and the inpatient hospitalization and death exclusions. Table 1 describes the prescriptions included in the analytic sample.

Table 1. Analytic sample selection (1/1/2014 – 11/30/2014)

	Total Medicaid-only beneficiaries (N)	Total dual-eligible beneficiaries (N)	Total Medicaid-only and dual-eligible beneficiaries age 18 years and older as of January 1, 2014 (N)	Total antipsychotic prescriptions (N)	Total beneficiaries with any antipsychotic prescriptions (N)	Total new antipsychotic prescriptions[a] (N)	Total new antipsychotic prescriptions, with exclusions[b] (N)
Level of analysis	**Beneficiary**	**Beneficiary**	**Beneficiary**	**Prescription**	**Beneficiary**	**Prescription**	**Prescription**
Total	**19,380,564**	**3,571,243**	**12,936,390**	**3,768,880**	**550,842**	**332,736**	**267,831**
Arkansas	771,712	137,575	435,175	79,666	13,163	7,918	5,751
Connecticut	718,183	181,383	583,689	222,802	29,609	17,266	13,150
Georgia	1,757,408	332,780	844,614	230,905	37,096	22,919	18,705
Iowa	611,524	93,854	391,053	132,056	18,277	11,247	8,705
Michigan	2,269,049	318,934	1,439,552	412,981	60,445	36,099	29,545
Mississippi	957,709	246,473	612,290	275,446	38,505	23,063	18,025
Missouri	615,772	172,184	371,026	102,246	17,345	10,392	8,640
New Jersey	1,559,258	242,949	1,038,323	274,640	38,217	22,704	17,937
New York	5,896,113	918,391	4,462,679	1,090,381	156,671	92,636	75,644
Pennsylvania	2,110,460	475,951	1,396,822	599,029	84,835	52,128	41,979
South Dakota	124,816	22,826	55,710	25,877	2,977	1,517	1,141
Tennessee	1,218,466	283,765	746,666	204,438	35,158	22,538	19,269
Vermont	168,349	37,646	136,889	33,391	4,948	3,201	2,525
West Virginia	529,276	94,121	390,341	76,330	12,176	8,164	6,083
Wyoming	72,469	12,411	31,561	8,692	1,420	944	732
Mean	1,292,038	238,083	862,426	251,259	36,723	22,182	17,855

Source: Mathematica analysis of 2013–2014 MAX PS, RX, OT, and IP files and Medicare Part B, Part D, Outpatient Department, and MedPAR files.

[a] An antipsychotic prescription is considered new if the beneficiary was not prescribed any antipsychotic medications within the previous 16 weeks.

[b] Exclusions limit the sample to prescriptions to beneficiaries who (1) were continuously enrolled in Medicaid for 16 weeks before and 4 weeks following the prescription, (2) did not die within 4 weeks following the prescription, and (3) did not have an inpatient stay within 4 weeks of prescription.

Slightly more than half of the prescriptions in the analytic sample were for dual-eligible beneficiaries. Almost a quarter of prescriptions in the analytic sample were filled for beneficiaries ages 18 to 34 (Table 2), whereas only 8.6 percent of prescriptions were filled for beneficiaries ages 65 to 74. Well over half of the prescriptions were for female beneficiaries (58.5 percent). White beneficiaries accounted for more than half of the new antipsychotic prescriptions included in the measure (57.0 percent), followed by black and Hispanic beneficiaries (26.5 and 10.4 percent, respectively).

Table 2. Descriptive statistics of the analytic sample

Characteristic	Number of qualifying prescriptions	Percentage
Total	**267,831**	**100.00**
Medicaid beneficiary category		
Medicaid only	132,835	49.60
Dual-eligible	134,996	50.40
Age		
18–34	65,162	24.33
35–44	45,932	17.15
45–54	59,976	22.39
55–64	44,690	16.69
65–74	23,067	8.61
75+	29,004	10.83
Sex		
Female	156,750	58.53
Male	111,081	41.47
Race/ethnicity		
White	152,748	57.03
Black	70,886	26.47
Hispanic	27,838	10.39
Other or unknown race/ethnicity	16,359	6.11
Disability status		
Disabled	180,564	67.42
Not disabled	87,267	32.58

Source: Mathematica analysis of 2013–2014 MAX PS, RX, and IP files and Medicare Part B, Part D, and MedPAR files. Disability status is determined by either Medicaid or Medicare enrollment data indicating disability. For all beneficiaries, we used the monthly Maintenance Assistance Status/Basis of Eligibility (MASBOE) variables in Medicaid enrollment data to determine disability in the month of the new antipsychotic prescription. For dual-eligible beneficiaries, we also used the Original Reason for Entitlement in the Medicare data.

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.

To analyze the impact of the exclusions on measure performance, we computed the denominator, numerator, and state-level performance for variations on the overall sample:

- Exclusions analysis 1: Removing the exclusion rule for beneficiaries who died during the four-week follow-up period

- Exclusions analysis 2: Removing the exclusion rule for beneficiaries who had an inpatient hospital stay during the four-week follow-up period
- Exclusions analysis 3: Removing both exclusion rules

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

As described in section 1.6, we collected information on the following variables using data extracted from Medicaid Analytic eXtract (MAX) 2013 and 2014 files: Medicaid eligibility category, age, sex, and race/ethnicity. This measure is based on a process that should be carried out for all beneficiaries (except those excluded), so no adjustment for patient mix is necessary. We did collect information about these variables and assessed disparities in performance rate for each group. Those results are described in section 2b5.

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*)

Signal-to-noise reliability. The signal-to-noise ratio (SNR) statistic, *R* (ranging from 0 to 1), summarizes the proportion of the variation between-state scores that is a result of real differences in the underlying characteristics of a state (such as differences in population demographic characteristics or quality of care provided) as opposed to background-level or random variation (such as measurement or sampling error). If *R*=0, there is no variation in the measure across entities, and all observed variation is due to sampling variation. In this case, the measure would not be useful for distinguishing between entities with respect to health care quality. Conversely, if *R*=1, all state scores are free of sampling error, and all variation represents real differences between entities in the measure result.

We estimated SNR reliability for the PMH-1 measure by using a beta-binomial model, which is suitable for binary pass/fail rate measures (Adams 2009). For PMH-1, the pass/fail designation is defined as the presence or absence of an eligible follow-up visit within a specified time frame after the prescription was filled. The beta-binomial model assumes that the state SNR score is a binomial random variable conditional on the state's true value, which comes from the beta distribution (ranging from 0 to 1). We calculated SNR reliability in three steps (Adams 2009; Adams 2014; NQF 2016).

First, we calculated state-specific PMH-1 variance ("noise") as a function of the measure "passing rate" at the state level, $\hat{p}$ (passed/eligible) and the sample size, *n*:

$$\sigma^2_{within} = \frac{\hat{p}(1-\hat{p})}{n}$$

Second, we used version 2.2 of the BETABIN SAS macro written by Wakeling (no date) to fit the beta-binomial model to the PMH-1 dataset. The macro produced the estimated average pass rate across all providers as well as the Alpha (α) and Beta (β) parameters that describe the shape of the fitted beta-binomial distribution. We calculated the "signal" (between-state variation of the PMH-1 measure) by using these parameters as follows:

$$\sigma^2_{between} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

Third, we calculated the SNR reliability as the ratio of the between-state variance and the total variance (i.e., the sum of the between-state and within-state variances) of the PMH-1 measure rate:

$$SNR = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

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*)

SNR analysis. The SNR statistic for PMH-1 was computed separately for each of the 15 states in the sample (Table 3). PMH-1 was highly reliable in distinguishing performance between states; the average SNR value was 0.98, and it ranged from 0.88 to 0.99 across states. It is important to note that although high reliability is not indicative of high quality health care, it does indicate that the measure may be used to distinguish between states with respect to health care quality.

Table 3. Signal-to-noise reliability for the PMH-1 measure (N = 15)

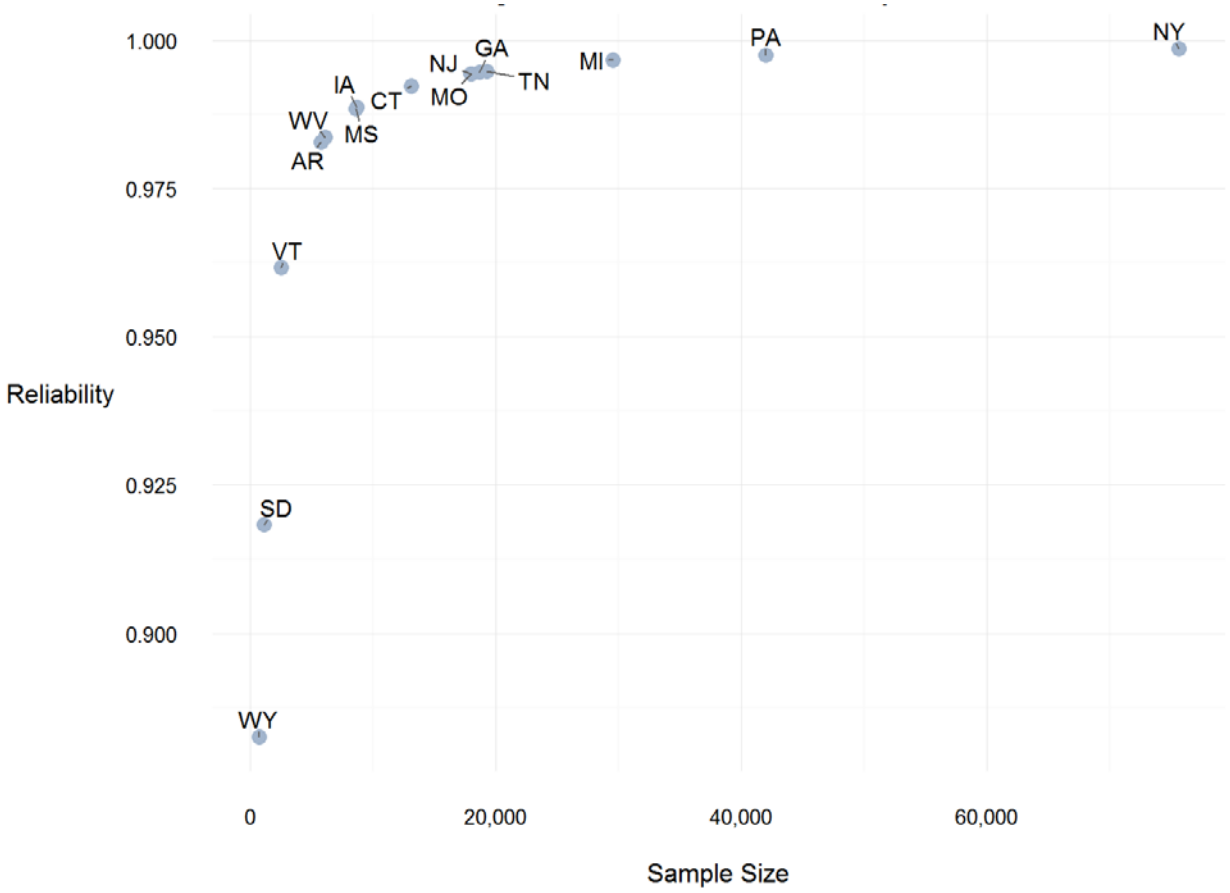
State	Number of new antipsychotic prescriptions (Denominator)	Number of prescriptions with follow-up visit within four weeks (Numerator)	Reliability score
Mean (all states)	17,855	8,719	0.98
Arkansas	5,751	2,558	0.98
Connecticut	13,150	6,234	0.99
Georgia	18,705	8,473	0.99
Iowa	8,705	4,986	0.99
Michigan	29,545	17,177	0.99
Mississippi	18,025	8,883	0.99
Missouri	8,640	3,924	0.99
New Jersey	17,937	8,420	0.99
New York	75,644	36,219	0.99
Pennsylvania	41,979	19,597	0.99
South Dakota	1,141	582	0.92
Tennessee	19,269	8,666	0.99
Vermont	2,525	1,401	0.96
West Virginia	6,083	3,225	0.98
Wyoming	732	440	0.88

Source: Mathematica analysis of 2013–2014 MAX PS, RX, OT, and IP files and Medicare Part B, Part D, OPD, and MedPAR files.

Note: The upper boundaries of the SNR statistic for MI, PA, and NY were truncated to 0.99 rather than rounded to 1.00 to reflect the uncertainty in the estimate.

High reliability for PMH-1 is likely supported by large enough sample sizes at the state level. The average number of newly filled prescriptions per state for PMH-1 was 17,855 (ranging from 259 to 42,227), and the average number of prescriptions with continuity treatment per state was 8,719 (ranging from 168 to 21,986). In Figure 1, we show the SNR statistics for each state plotted against sample size and demonstrate how reliability increases with sample size. The reliability of all states was above the NQF threshold of 0.70 for acceptable reliability (Measure Testing Task Force Report 2011).

Figure 1. State-level signal-to-noise ratio reliability of PMH-1 rates



Source: Mathematica analysis of 2013–2014 MAX PS, RX, OT, and IP files and Medicare Part B, Part D, OPD, and MedPAR files.

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?*)

PMH-1 is rated high for scientific acceptability, based on reliability results. The SNR analyses showed that the reliability of PMH-1 is good, and the high SNR indicates the measure performance results are highly precise with a very low degree of measurement error.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

Face and convergent validity testing was conducted using state-level performance scores.

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

Face validity. To assess face validity, we conducted a survey of TEP members to obtain their assessment of the extent to which the measure's state-level performance scores distinguish good quality from poor quality of care. Of the 11 TEP members who responded to the survey, eight TEP members agreed that the measure distinguishes good quality from poor quality of care. Two TEP members reported a neutral assessment of the performance score's face validity, and one TEP member disagreed with the statement. Overall, this support suggests that the measure has moderate-high face validity for distinguishing the quality of care provided to beneficiaries who are newly prescribed antipsychotic medications.

Convergent validity. To test convergent validity, we examined how state-level performance of PMH-1 compares to state-level performance on several other measures of similar concepts, based on input from TEP members. We used the following criteria to identify relevant measures for inclusion in this additional analysis:

1. The measure shares an underlying mechanism of care coordination or follow-up with PMH-1, is related to behavioral health care or conditions, or is focused on medication adherence or management
2. The measure performance rates are publicly available at the state level in FFY2015 (the period corresponding with our PMH-1 testing data)
3. The measure performance rates are available for the majority of states in our PMH-1 testing report

Five Core Set measures met these criteria and were included in this analysis:

- Adherence to Antipsychotics for Individuals with Schizophrenia (SAA-AD)
- Annual Monitoring for Patients on Persistent Medications (MPM-AD)
- Antidepressant Medication Management (AMM-AD)
- Follow-up After Hospitalization for Mental Illness (FUH-AD)
- Follow-up Care for Children Prescribed Attention-Deficit/Hyperactivity Disorder (ADHD) Medication (ADD-CH)

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

For the state-level convergent validity analysis, the PMH team created comparative graphs for visual inspection that compares state-level performance on PMH-1 to performance on each of the five Core Set measures, respectively. A comparative graph provides suggestive evidence of PMH-1's convergent validity if high performance on PMH-1 relative to other states was correlated with high performance on the Core Set measures relative to other states, and low performance on PMH-1 relative to other states was correlated with low performance on the Core Set measures relative to other states. The sample of states was too small to conduct these analyses formally (that is, statistically).

We find that states with high performance (that is, relatively close to 60.1 percent, the highest state-level performance rate observed during testing) on PMH 1 often had relatively high performance on the following measures:

- Follow-up After Hospitalization for Mental Illness
- Antidepressant Medication Management
- Follow-up Care for Children Prescribed ADHD Medications

The reverse was true as well—states with low performance (that is, relatively close to 45.5 percent, the lowest state-level performance rate observed during testing) on PMH-1 also had low performance on these three Core Set measures.

Table 4 shows states performance on Follow-up After Hospitalization for Mental Illness (FUH-AD) from highest to lowest, and Figure 2 compares states' performance on PMH-1 to performance on FUH-AD. State performance on FUH-AD ranged from 45.6 percent in Iowa to 75.9 percent in Tennessee. Based on the ten states with available data, we found that most states exhibit a roughly positive correlation between the two measures. For example, Missouri has low relative performance on both—45.4 percent on PMH-1 and 50.3 percent on FUH-AD, but Vermont has relatively high

performance on both—55.5 percent on PMH-1 and 75.4 percent on FUH-AD. Tennessee and Iowa are the exceptions, the latter of which is an outlier in almost all cases.

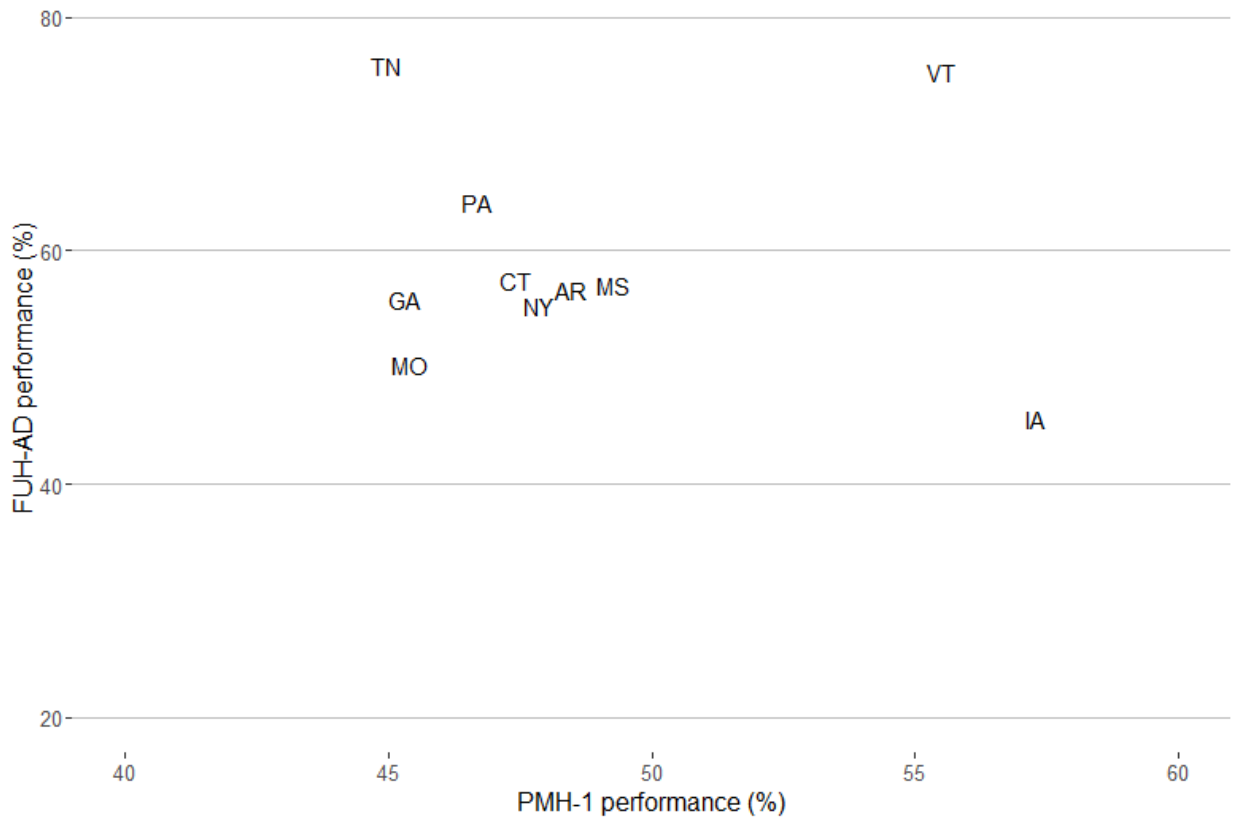
Table 4. State performance on Follow-up After Hospitalization for Mental Illness (FUH-AD)

State	FUH-AD performance (%)
Tennessee	75.9
Vermont	75.4
Pennsylvania	64.2
Connecticut	57.5
Mississippi	57.1
Arkansas	56.7
Georgia	55.9
New York	55.3
Missouri	50.3
Iowa	45.6

Source: State-reported performance rates for Adult Core Set measures, FFY2015.

Note: Michigan, New Jersey, South Dakota, West Virginia, and Wyoming did not report state performance rates for the FUH-AD measure.

Figure 2. State Performance on PMH-1 and Follow-up After Hospitalization for Mental Illness (FUH-AD)



Source: Mathematica analysis of state-level performance on Adult Core Set measures, FFY2015 and state-level performance of PMH-1 using data from 2013–2014 MAX PS, RX, OT, and IP files, and Medicare Part B, Part D, OPD, and MedPAR files.

Table 5 shows state performance on Antidepressant Medication Management (AMM-AD) from highest to lowest, and Figure 3 compares states' performance on PMH-1 to performance on AMM-AD. State performance on AMM-AD ranges

from 33.3 percent in Arkansas to 69.3 percent in Vermont, and most of the states appear clustered around 50 percent performance on AMM-AD and about 48 percent performance on PMH-1. While Iowa again appears to be somewhat of an outlier, states with higher relative performance on PMH-1 often have higher relative performance on AMM-AD.

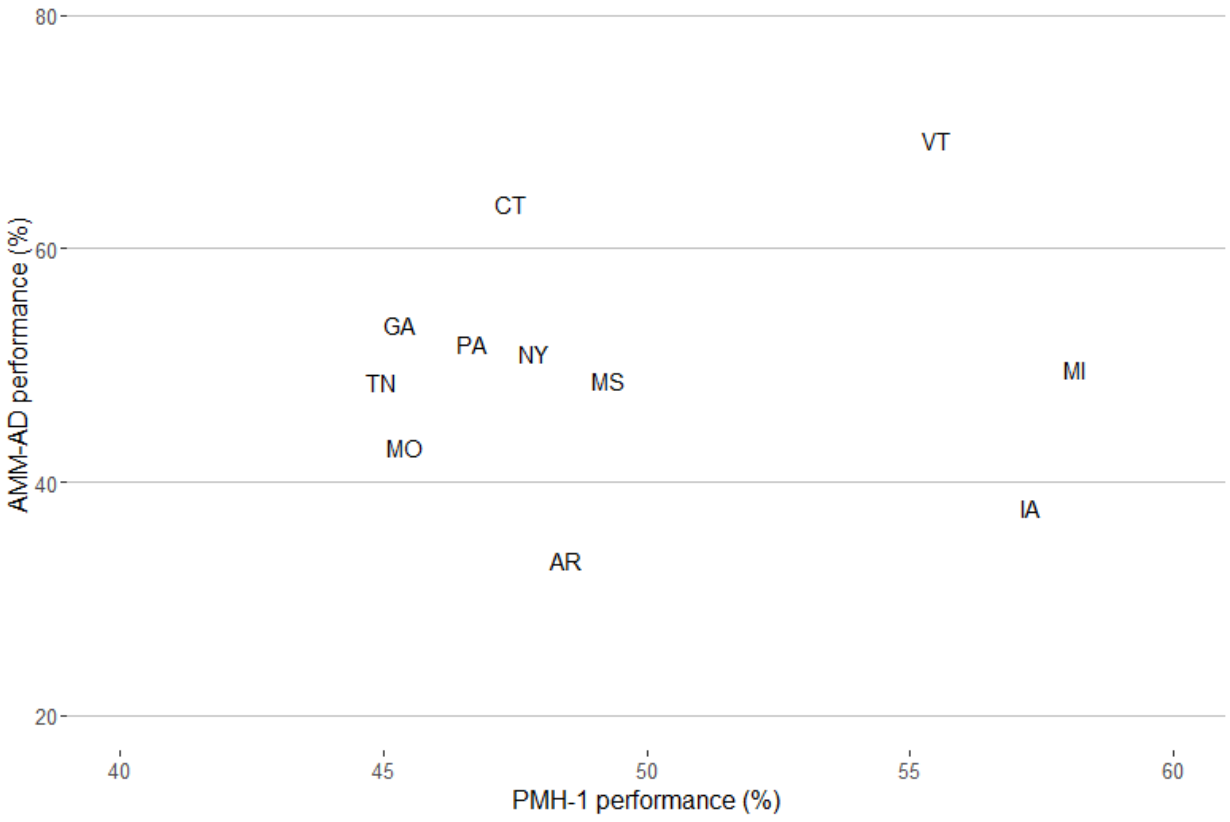
Table 5. State performance on Antidepressant Medication Management (AMM-AD)

State	AMM-AD performance (%)
Vermont	69.3
Connecticut	63.9
Georgia	53.5
Pennsylvania	51.9
New York	51.0
Michigan	49.7
Mississippi	48.8
Tennessee	48.6
Missouri	43.0
Iowa	37.9
Arkansas	33.3

Source: State-reported performance rates for Adult Core Set measures, FFY2015.

Note: New Jersey, South Dakota, West Virginia, and Wyoming did not report state performance rates for the AMM-AD measure.

Figure 3. State Performance on PMH-1 and Antidepressant Medication Management (AMM-AD)



Source: Mathematica analysis of state-level performance on Adult Core Set measures, FFY2015 and state-level performance of PMH-1 using data from 2013–2014 MAX PS, RX, OT, and IP files, and Medicare Part B, Part D, OPD, and MedPAR files.

Table 6 shows state performance on Follow-up Care for Children Prescribed ADHD Medication (ADD-CH) from highest to lowest, and Figure 4 compares states' performance on PMH-1 to (ADD-CH). ADD-CH ranges from 27.9 percent in Pennsylvania to 66.7 percent in Vermont. Although Michigan and Iowa are again moderate outliers, there appears to be a somewhat positive correlation between PMH-1 and ADD-AD. Vermont has high relative performance on both measures—55.5 percent on PMH-1 and 66.7 percent on ADD-CH. By contrast, Pennsylvania has low relative performance on both measures – 46.7 percent on PMH-1 and 27.9 percent on ADD-CH.

Table 6. State performance on Follow-up Care for Children Prescribed ADHD Medication (ADD-CH)

State	ADD-CH performance (%)
Vermont	66.7
Arkansas	61.6
New York	57.8
Connecticut	57.2
Mississippi	56
West Virginia	50.6
Tennessee	47.8
Michigan	38.9
Georgia	35.5
Iowa	32.7
New Jersey	32.5
Pennsylvania	27.9

Source: State-reported performance rates for Child Core Set measures, FFY2015.

Note: Missouri, South Dakota, and Wyoming did not report state performance rates for the ADD-CH measure

Figure 4. State Performance on PMH-1 and Follow-up Care for Children Prescribed ADHD Medication (ADD-CH)

Source: Mathematica analysis of state-level performance on Child Core Set measures, FFY2015 and state-level performance of PMH-1 using data from 2013–2014 MAX PS, RX, OT, and IP files, and Medicare Part B, Part D, OPD, and MedPAR files.

We did not find a relationship in state performance between PMH-1 and Adherence to Antipsychotics for Individuals with Schizophrenia (SAA-AD) or Annual Monitoring for Patients on Persistent Medications (MPM-AD). There was very little variation in state performance on these measures, which could explain these findings. State level SAA-AD measure performance ranged from 59.2 percent in Arkansas to 71.7 percent in Pennsylvania (to provide context, performance on the AMM-AD measure ranged from 33.3 percent in Arkansas to 69.3 percent in Vermont). There was even less variation by state on the Medication Monitoring measure (MPM-AD); all states had very high performance (over 87 percent). Where there is little state variation in performance on a Core Set measure, there is minimal leverage with which to compare it to PMH-1, and therefore difficult to determine whether there is a correlation between the two.

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

PMH-1 is rated moderate for validity. Most respondents from the project's TEP assessed the measure's performance scores as able to distinguish between good quality and poor quality of care for Medicaid beneficiaries newly prescribed antipsychotic medications. Additionally, state-level PMH-1 rates demonstrated some association with several related state-level rates of measures of similar concepts.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b3

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

We tested the impact of the two exclusion rules to the PMH-1 measure rates proposed by the Clinical Advisory Workgroup:

1. **Hospitalization** - Medicaid beneficiaries with an acute inpatient admission during the four-week follow-up period after prescription of an antipsychotic medication; and
2. **Death** - Patients who expired within four weeks of new prescription date.

The current PMH-1 definition excludes both cases in the denominator definition. Based on this, we tested the measure rate in the following four scenarios:

1. PMH-1 measure specification (pmh1_current): the current measure specification (i.e., exclude both cases from denominator)
2. Exclusions analysis 1 (pmh1_nodeath): do not apply the death exclusion rule (i.e., only exclude those beneficiaries with an acute inpatient admission during the four-week follow-up period after prescription of an antipsychotic medication)
3. Exclusions analysis 2 (pmh1_nohosp): do not apply the hospitalization exclusion rule (i.e., only exclude those who died within four weeks of new prescription date)
4. Exclusions analysis 3 (pmh1_noexcld): neither exclusion rule applied

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Denominator sizes. The exclusion rule has small influence on the eligible denominator size. The measure with both exclusion rules applied (i.e., current measure specs) includes 94.7 percent (or 267831 out of 282905) of the denominator as oppose to the measure not applying any exclusion rule. In general, the hospitalization rule excluded more prescriptions (95.2 percent) from the denominator than the death rule (99.4 percent). This was the case across all states

other than New York, for which the death rule excluded more prescriptions from the denominator than the hospital rule.

Measure performance. Removing either or both of the hospitalization and death exclusion rules result in a slight change on the overall PMH-1 measure rates (Table 7). Moreover, variations on measure results are small across all states, and performance rankings are (almost) the same between all pairs, as evidenced by Spearman rank correlations extremely close or equal to 1 (Table 8).

Table 7. PMH-1 performance rates (all states), with and without exclusions applied

Exclusions applied	PMH-1 rate (all states)
Exclude if inpatient hospitalization or died during follow-up period	48.8%
Exclude if inpatient hospitalization during follow-up period	48.7%
Exclude if died during follow-up period	49.1%
No exclusions	49.0%

Source: Mathematica analysis of 2013–2014 MAX PS, RX, and IP files and Medicare Part B, Part D, and MedPAR files.

Table 8. Spearman rank correlations for PMH-1 rates with and without exclusions applied

	pmh1_nodeath	pmh1_nohosp	pmh1_noexcld	pmh1_current
pmh1_nodeath	1	.997	.997	1
pmh1_nohosp	.997	1	1	.997
pmh1_noexcld	.997	1	1	.997
pmh1_current	1	.997	.997	1

Source: Mathematica analysis of 2013–2014 MAX PS, RX, and IP files and Medicare Part B, Part D, and MedPAR files.

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Although the exclusions have limited impact on the denominator size and measure performance, we maintained the exclusions in the measure specifications, given that (1) the clinical advisory workgroup contributed to the development of these exclusions and the Technical Expert Panel supported the exclusions as appropriate during their review of testing results, and (2) the exclusions are relatively straightforward to calculate and are not expected to impose significant additional burden in measure implementation.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with _risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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

Not applicable

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

Not applicable

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

Not applicable

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

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

Not applicable

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

Not applicable

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

Not applicable

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

Not applicable

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

Not applicable

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

Not applicable

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

Not applicable

2b3.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

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

To assess whether performance differed meaningfully among states included in testing, we examined the distribution of measure performance (for example, mean, median, minimum, 25th percentile, 75th percentile, and maximum) across states. In addition, we calculated the 95 percent confidence interval of the measure rate for each state using a z-distribution for proportion. We then compared each state's confidence interval to the overall measure rate that uses all beneficiaries across states. States with measure rates significantly lower than the overall rate indicate evidence of less-than-optimal performance, which suggests room for improvement.

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

We found that measure rates across the fifteen states covered a wide range with meaningful variation. Specifically, the measure rate ranged from 44.5 percent to 60.1 percent with a mean of 50.2 percent and standard deviation of 5.29 percent. When looking into state-specific measure rate, eight of the 15 states (53.3 percent) exhibited significantly lower measure rates than the average performance, with their 95 percent confidence intervals under the overall performance rate.

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

These findings suggest room for improvement in follow-up after prescription of new antipsychotic medications in the states included in testing. Five states showed significantly higher measure rates than the average performance, and two states had performance statistically not distinguishable from the average performance. Overall, these findings indicate both statistically significant and practically meaningful differences in PMH-1 performance. Interpretation of these comparisons should be tempered by the fact that only 15 states were included in this analysis; the mean rate for the entire nation could be different. Moreover, when discussing room for improvement, states that are not statistically different from or even above the mean rate of 50.2 percent have room for improvement. Only one state reached a rate of 60 percent for follow-up care after a new antipsychotic prescription.

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

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

Not applicable; only one set of specifications.

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

demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

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

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

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

_____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

The extent of missing data was assessed using the MAX validation and anomaly tables (cited below).

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

The vast majority of the data elements required to calculate the measure—dates of service, date of birth, Medicaid eligibility, prescription fill date, National Drug Code (NDC), and type of service—have negligible missingness in MAX data for the states in the study sample. The one data element that is missing information for more than 5 percent of claims is the field for CPT/HCPCS code.

Table 9 below contains eligibility-related missingness information from 2013 and where available, 2014 MAX data. Overall, well over 95 percent of MAX claims can be matched to eligibility information.

Table 9. Percent of claims missing corresponding Medicaid eligibility information

State	Year	% with Claims and Missing Medicaid Eligibility (Excludes S-CHIP Only)	IP: % Missing Eligibility and > $0 Paid (Excludes S-CHIP Only)	LT: % Missing Eligibility and > $0 Paid (Excludes S-CHIP Only)	OT: % Missing Eligibility and > $0 Paid (Excludes S-CHIP Only)
AR	2013	2.46	5.33	0.29	0.44
CT	2013	0.27	0.22	0.07	0.18
GA	2013	0.96	0.12	0.02	0.17
	2014	0.85	0.07	0.01	0.16
IA	2013	0.17	0.14	0.04	0.01
	2014	0.08	0.04	0.02	0.00
MI	2013	4.08	1.59	0.41	0.38
	2014	1.50	0.94	0.46	0.10
MO	2013	2.03	0.14	0.01	0.97
	2014	0.35	0.21	0.02	0.07
MS	2013	0.11	0.54	0.02	0.04
	2014	0.23	0.28	0.02	0.10
NJ	2013	0.55	0.20	0.42	0.21
	2014	0.55	0.24	0.33	0.21
NY	2013	0.07	0.23	0.21	0.00
PA	2013	2.82	1.01	0.47	0.08
	2014	3.77	0.94	0.16	0.31
SD	2013	0.01	0.00	0.00	0.01
	2014	0.01	0.00	0.00	0.00
TN	2013	0.39	0.00	0.00	0.03
	2014	0.56	0.00	0.00	0.03
VT	2013	0.53	0.93	0.44	0.17
	2014	0.22	0.41	0.29	0.04
WV	2013	3.60	0.14	0.01	0.19
	2014	0.09	0.05	0.02	0.01
WY	2013	0.57	1.36	0.23	0.20
	2014	0.75	1.65	0.32	0.18

Source: MAX validation tables. Available at the following URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports.html?DLSort=0&DLEntries=10&DLPage=1&DLSortDir=ascending

Table 10 below contains missingness related to claims data elements used in the calculation of the measure.

Table 10. Percent of FFS claims missing data elements used in the calculation of the measure

State	Year	% IP Stays (MAX TOS=01)	OT % with HCPCS or CPT-4 Code
Arkansas	2013	100.0	100.0
Connecticut	2013	98.6	82.3
Georgia	2013	100.0	100.0
	2014	100.0	100.0
Iowa	2013	100.0	94.9
	2014	100.0	94.3
Michigan	2013	100.0	95.4
	2014	100.0	93.1
Mississippi	2013	100.0	100.0
	2014	100.0	100.0
Missouri	2013	99.3	100.0
	2014	99.3	100.0
New Jersey	2013	99.2	100.0
New York	2013	100.0	26.3
Pennsylvania	2013	99.6	100.0
	2014	99.6	100.0
South Dakota	2013	100.0	99.9
	2014	100	99.9
Tennessee	2013	0.0	100.0
	2014	0.0	100.0
Vermont	2013	99.8	100.0
	2014	99.9	100.0
West Virginia	2013	100.0	100.0
	2014	100.0	100.0
Wyoming	2013	97.7	100.0
	2014	97.7	100.0

Source: MAX anomaly tables. Available at the following URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html.

Note: Tennessee had virtually no inpatient FFS claims in 2013 or 2014.

There are two states that have more than 10 percent of other/professional (OT) claims missing HCPCS or CPT code—Connecticut and New York. New York frequently uses state-specific CPT/HCPCS codes, although their use may be restricted to particular kinds of services. Table 10 displays the percent of claims missing national HCPCS/CPT codes and cannot distinguish between completely missing codes and the use of state-specific codes (both of which would be considered "missing"). CPT codes are important for identifying follow-up visits. Some follow-up visits may be missed, particularly in New York, and could result in lower measure performance for that state.

Type of service is necessary for identifying whether the beneficiary had an inpatient, prescription, other service, or SNF stay, and as a proxy for prescribing authority. This data element is virtually never missing in MAX data because it is used for binning claims into different files (inpatient, prescription, other services, etc.); see, for example in Table 10 – almost every FFS inpatient (IP) claim has a type of service of "01" ("inpatient"). The one exception is Tennessee, which has virtually no inpatient FFS claims in 2013 or 2014.

The measure hinges on identifying beneficiaries taking a new antipsychotic prescription during the measurement year, which is identified using NDC and prescription fill date. In no states are claims missing NDC. Further, prescription fill date

is a required field to be included in MAX data and therefore would not be missing for any claims. Similarly, dates of service are required fields for MAX IP and OT claims and therefore are not missing.

Finally, Table 11 shows missingness of date of birth, sex, or race, which we used to compute measure performance by subgroups. A very low percentage of Medicaid enrollees have missing date of birth or sex. There are relatively high levels of enrollees missing race, so tabulations of measure performance by race will only be possible for a subset of the population. Date of death is only populated for beneficiaries who died, so it will be missing—by design—in most cases.

Table 11. Percent of Medicaid enrollees with missing date of birth, sex, or race

State	MAX year	Percent of enrollees missing date of birth	Percent of enrollees with missing sex	Percent of enrollees with missing race
Arkansas	2012	0.0	0.0	14.6
Connecticut	2012	0.0	0.0	0.0
Georgia	2013	0.0	0.0	11.1
Iowa	2013	0.0	0.0	43.8
Michigan	2012	0.0	0.0	10.7
Mississippi	2013	0.0	0.0	6.1
Missouri	2012	0.0	0.0	4.1
New Jersey	2012	0.0	0.0	28.0
New York	2013	1.3	1.0	7.7
Pennsylvania	2013	0.0	0.0	12.3
South Dakota	2013	0.0	0.0	0.0
Tennessee	2013	0.0	0.0	10.9
Vermont	2013	0.0	0.0	26.2
West Virginia	2013	0.0	0.0	1.5
Wyoming	2013	0.0	0.0	14.4

Source: MAX anomaly tables. Available at the following URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html.

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

Given the relatively small amount of missing information, we don't believe there is any systematic bias. In addition, states implementing the measure will likely have even less missing data because they will be able to account for their state-specific codes when constructing the measure.

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.

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

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*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic claims

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Not applicable.

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)*.

Not applicable, no fees or licensing are currently required

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.

Specific Plan for Use	Current Use (for current use provide URL)
Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

Not applicable; the measure is under initial endorsement review and is not currently used in an accountability program.

4a1.2. 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?*)

CMS is considering implementation options for this measure. There are no identified barriers to implementation in a publicly reporting or accountability application.

4a1.3. 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.*)

CMS is developing measures to improve the quality of care of the following Medicaid populations served by CMS's Innovation Accelerator Program:

- People eligible for both Medicare and Medicaid, or "Dual-eligible beneficiaries"
- People receiving long-term services and supports (LTSS) through managed care organizations
- People with substance use disorders; beneficiaries with complex care needs and high costs; beneficiaries with physical and mental health needs; or Medicaid beneficiaries who receive LTSS in the community

This measure is intended for voluntary use by states to monitor and improve the quality of care provided for Medicaid beneficiaries with physical and mental health integration needs. States may choose to begin implementing the measures based on their programmatic needs.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Not applicable.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Not applicable.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Not applicable.

4a2.2.2. Summarize the feedback obtained from those being measured.

Not applicable.

4a2.2.3. Summarize the feedback obtained from other users

Not applicable.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Not applicable.

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

This measure is being considered for initial endorsement. Adoption of this performance measure has the potential to improve the quality of care for Medicaid beneficiaries who are newly prescribed antipsychotic medications. Currently, on average, states are providing follow-up care within four week for only about half of new antipsychotic prescriptions (50.2 percent). These findings suggest room for improvement in follow-up care in the states included in testing, which can be important for monitoring side effects, providing patient education, and adjusting dosage or medications as needed. Even states that had relatively high performance on the measure likely have room for improvement, as only one state reached a rate of 60 percent for follow-up care after a new antipsychotic prescription. This measure will encourage accountable entities to ensure that beneficiaries receive a timely follow-up visit to monitor their new antipsychotic medication. Initiatives to promote improvement could include deployment of care management or coordination strategies to engage participants in follow-up care, as well as an increased focus on communication strategies between providers for beneficiaries who are expected to receive follow-up care in settings different from where they were prescribed the initial antipsychotic medication.

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

Not applicable. This measure has not been implemented yet. There were no unexpected findings identified during testing of this measure.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

Not applicable. This measure has not been implemented yet. There were no unexpected findings identified benefits during testing of this measure.

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.

Yes

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

0108 : Follow-Up Care for Children Prescribed ADHD Medication (ADD)

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

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

Yes

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

This measure differs from NQF 0108 in that it focuses on adults rather children, and on antipsychotic medications rather than ADHD medications. The measures are completely harmonized to the extent possible, with the same follow-up period and look-back period to establish a "new prescription."

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.)

Not applicable.

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.

No appendix Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare and Medicaid Services (CMS)

Co.2 Point of Contact: Roxanne, Dupert-Frank, Roxanne.Dupert-Frank@cms.hhs.gov, 410-786-9667-

Co.3 Measure Developer if different from Measure Steward: Mathematica Policy Research

Co.4 Point of Contact: Henry, Ireys, hireys@mathematica-mpr.com, 630-792-5073-

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 project team convened a Clinical Advisory Workgroup to consult on the development of measure specifications, including the appropriate follow-up period after a new prescription, the appropriate look-back period to determine that a prescription should be considered a "new" prescription, and the appropriate visits to include in the numerator. The Clinical Advisory Workgroup included the following members:

- Robert Cotes, MD (American Psychiatric Association)
- Thomas Croghan, MD (Mathematica Policy Research)
- Catherine Fullerton, MD, MPH (Truven Health Analytics)
- Aaron Garman, MD (American Academy of Family Physicians)
- David Mancuso, PhD (Washington State Department of Social and Health Services)
- Robert McCarron, DO (American College of Physicians)
- Benjamin Miller, PsyD (University of Colorado School of Medicine)
- Clifford Moy, MD (TMF Health Quality Institute)
- Joseph Parks, MD (Missouri Department of Social Services)
- Cindy Thomas, PhD (Brandeis University)
- Keith Widmer, RPh, BCPP (Pharmacy Quality Alliance)
- Amy Windham, PhD, MPH (Truven Health Analytics)
- Barbara Zarowitz, PharmD (Pharmacy Quality Alliance)

The project's Technical Expert Panel provided in put on measure selection, feedback on testing results, and an assessment of the face validity of performance scores. The TEP includes the following members:

Consumer Representative 1

-Carol McDaid (Capitol Decisions, Inc)

Consumer Representative 2

-Janice Tufte (Patient-Centered Outcomes Research Institute (PCORI) ambassador)

Consumer Representative 3

-Kayte Thomas (PCORI ambassador)

State Official 1

-Joe Parks (Missouri HealthNet Division (Medicaid))

State Official 2

-David Mancuso (Washington State Department of Social and Health Services)

State Official 3

-Roxanne Kennedy (New Jersey Division of Mental Health and Addiction Services)

Health Plan Representative 1

-Alonzo White (Aetna Medicaid)

Health Plan Representative 2

-Deb Kilstein (Association for Community Affiliated Plans)

Health Plan Representative 3

-Jim Thatcher (Massachusetts Behavioral Health Partnership, Beacon Health Options)

Provider Organization Representative 1

-Daniel Bruns (Health Psychology Associates)

Provider Organization Representative 2

-Aaron Garman (Coal Country (ND) Community Health Center and American Academy of Family Practice Comm. on Quality & Practice)

Provider Organization Representative 3

-Annette DuBard (Community Care of North Carolina)

Subject Matter Expert/Researcher 1

-Andrew Bindman (University of California San Francisco)

Subject Matter Expert/Researcher 2

-Mady Chalk (Treatment Research Institute)

Subject Matter Expert/Researcher 3

-Kimberly Hepner (RAND Corporation)

Subject Matter Expert/Researcher 4

-Benjamin Miller (University of Colorado School of Public Health)

Subject Matter Expert/Researcher 5

-Alex Sox-Harris (Department of Veterans Affairs)

Federal Agency Official 1

-Deb Potter (Office of the Assistant Secretary for Planning and Evaluation)

Federal Agency Official 2

-Lisa Patton (Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality)

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? Specifications for this measures will be reviewed and updated annually.

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement: Limited proprietary coding is contained in the Measure specifications for user convenience. Users of proprietary code sets should obtain all necessary licenses from the owners of the code sets. Mathematica disclaims all liability for use or accuracy of any CPT or other codes contained in the specifications.

CPT(R) contained in the Measure specifications is copyright 2004-2016 American Medical Association.

Ad.7 Disclaimers: These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 Additional Information/Comments: Not applicable.

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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3315e

Measure Title: Use of Antipsychotics in Older Adults in the Inpatient Hospital Setting

Measure Steward: Centers for Medicare & Medicaid Services

Brief Description of Measure: Proportion of inpatient hospitalizations for patients 65 years of age and older who receive an order for antipsychotic medication therapy.

Developer Rationale: Clinical guidelines recommend against using antipsychotics as a standard first line of treatment for patients experiencing delirium or behavioral and psychological symptoms of dementia unless they present a threat to themselves or their caregivers (AGS 2015a, AGS 2015b, NICE 2016, Reus 2016). Antipsychotics are often used off-label as a method of treating patients in an acute confusional state despite conflicting evidence regarding the effectiveness of antipsychotics in treating these disorders (Neufeld 2016, Barr 2013, Campbell 2009). The benefits of this measure lie in the potential to reduce inappropriate use of antipsychotics in inpatient hospital settings and the unnecessary continuation of the intervention post-discharge, resulting in improved patient outcomes (reduced morbidity and mortality) for older adults. Measuring the use of antipsychotics among hospitalized older adult patients could help shift the focus to determining alternative causes of this behavior (such as medication interactions, medication side effects, etc.) and adjusting treatment accordingly.

American Geriatrics Society 2015 Beers Criteria Update Expert Panel. (2015a). American Geriatrics Society 2015 Updated Beers Criteria for Potentially Inappropriate Medication Use in Older Adults. J Am Geriatr Soc, 63(11), 2227-

AGS Expert Panel on Postoperative Delirium in Older Adults. "American Geriatrics Society abstracted clinical practice guideline for postoperative delirium in older adults." J Am Geriatr Soc, 63(1), 2015b, pp 142-50. doi: 10.1111/jgs.13281.

Barr, J., Fraser, G.L., Puntillo, K., et al. (2013). Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. Crit Care Med, 41(1), 263-306.

Campbell, N., Boustani, M.A., Ayub, A., et al. (2009). Pharmacological management of delirium in hospitalized adults--a systematic evidence review. J Gen Intern Med, 24(7), 848-53.

Neufeld, K.J., Yue, J., Robinson, T.N., et al. (2016). Antipsychotic Medication for Prevention and Treatment of Delirium in Hospitalized Adults: A Systematic Review and Meta-Analysis. J Am Geriatr Soc, 64(4), 705-714.

NICE (National Institute for Health and Clinical Excellence) Dementia: Supporting people with dementia and their carers in health and social care. 2016 (Issued November 2006, Modified September 2016).

Reus, V.I., Fochtmann, L.J., Eyler, A.E., et al. (2016). The American Psychiatric Association Practice Guideline on the Use of Antipsychotics to Treat Agitation or Psychosis in Patients with Dementia. Am J Psychiatry, 173(5), 543-546.

Numerator Statement: : Inpatient hospitalizations for patients who received an order for an antipsychotic medication during the inpatient encounter.

Denominator Statement: Denominator: Non-psychiatric inpatient hospitalizations for patients who are 65 and older.

Denominator Exclusions: Denominator Exclusions: Inpatient hospitalizations for patients with a diagnosis of schizophrenia, Tourette´s syndrome, bipolar disorder, Huntington´s disease during the encounter.

Measure Type: Process

Data Source: Electronic Health Records

Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date: Not applicable; this measure is not a paired or grouped measure.

Staff Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* 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. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- Logic model provided describing the steps between the healthcare structures and processes and patient's health outcome(s).
- Clinical Practice Guideline recommendations on the use of Antipsychotics:
    - American Geriatrics Society Guideline 2015 Beers Criteria for Potentially Inappropriate Medication Use in Older Adults recommendation to avoid antipsychotics (except for schizophrenia and bipolar disorder, or as short-term use as antiemetic during chemotherapy). **Moderate grade assigned to the evidence and Strong grade assigned to the recommendation** (e.g. the benefits clearly outweigh harms).
    - American Psychiatric Association Guideline on the Use of Antipsychotics to Treat Agitation or Psychosis in Patients with Dementia (2016). **Moderate grade assigned to the evidence** and **Recommendation grade assigned to the recommendation** (e.g. indicates confidence that the benefits clearly outweigh harms).

Exception to evidence
N/A

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*

- *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: moderate; Consistency: high (Box 5) → Moderate (Box 5b) → Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE:

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 rationale for this measure that include the potential benefit to reduce inappropriate use of antipsychotics in inpatient hospital settings and the unnecessary continuation of the intervention post-discharge, resulting in improved patient outcomes for older adults after discharge.

The developer provides performance results at the facility level from two health systems, which provided data from 10 hospitals, and one critical access hospital. The data were derived from three test sites using two different EHRs and representing 137,817 hospital encounters.

Overall summary statistics from all three sites:

Mean	Standard Deviation	Min	10th Percentile	Interquartile Range	90th Percentile	Max
15.3%	5.2%	5.5%	8.4%	5.9%	20.4%	22.8%

The developer cites additional research on the use of antipsychotics during inpatient hospital visits to support opportunity for improvement:

- A retrospective cohort study of roughly 18,000 adult non-psychiatric hospital admissions over a year found antipsychotic exposure in 9 percent of visits (Herzig 2016a).
- A retrospective cohort study of 2,700,000 adult non-psychiatric hospital admissions over a year found antipsychotic exposure in 6 percent of visits (Marshall 2016).

Disparities

The developer notes that the research on disparities in the use of antipsychotics is limited. They provide analysis of antipsychotic ordering rates on insurance coverage, gender and race:

- Medicare and Medicaid coverage had the highest rate of antipsychotic ordering
- Statistically significant differences in antipsychotic orders for gender and race

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are there other reasons to minimize the use of antipsychotics in inpatient setting other than decrease in continuation post-discharge?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence

Comments:

** Considerable evidence exists to suggest that antipsychotics must be used with great prudence in the geriatric population, particularly for individuals with cognitive impairment. However, there did not seem to be literature supporting that use of antipsychotic medication in a medical setting to address agitation/delirium is then continued upon discharge.

** Systematic reviews have been conducted with moderate grades assigned to evidence and that benefits outweigh harms. Potentially inappropriate use of antipsychotics without assessment of underlying conditions and diagnoses of serious mental illness, especially in older adults, can lead to increased mortality. Evidence supports creation of this measure. Expert opinion was generated leading to APA Guideline.

** This process measure is intended to look at use of antipsychotics in hospitalized elderly. The intention is to reduce the inappropriate use of this class of medications with elderly patients. There is evidence in the literature of inappropriate use.

** Important given the potential negative impact of anti-psychotics on older adults. Unclear how much will improve the evaluation and treatment of agitation and other symptoms due to delirium/dementia, whether in hospital, after discharge or in lower levels of care like ICF, SNFs.

** There is good evidence that the measure does support increased safety (decreased strokes and respiratory depression) in significant ways. I am convinced that it is important inpatient as outpatient. The one suggestion I have is to consider modifying it so that one could document and exclude cases where the patient has dementia and IS behaviorally a high risk to self or others.

** The evidence is tangential. The developers note that the APA recommends that nonemergency antipsychotic medication should only be used for treatment of agitation or psychosis in patients with dementia when symptoms are severe, are dangerous, or cause significant distress to the patients. These conditions are not considered as exclusions.

** Evidence provided by developers indicates off label use of antipsychotics for hospitalized non-psychiatric patients has been associated with higher rates of cerebrovascular accidents, inducing or worsening delirium, and other CNS adverse events. Evidence includes 8 systematic reviews, 5 randomized control studies, and 14 cohort studies. There appears to be a relatively strong case to argue against the use of antipsychotic medication for older persons experiencing a non-psychiatric inpatient hospitalization.

** The evidence comes from AGS and APA guidelines. In the former, the evidence is Moderate due to the number, size, consistency and generalizability which limit the strength. The evidence does apply directly and relates to the desired outcome.

** The evidence is moderate but leans towards the overreliance on antipsychotic use in elderly patients resulting in strokes/morbidity.

** Early testing of this measure seems to accurately capture that medications are ordered 5.5-19.4% of the time depending on size of hospital sample. I don't know what to compare this number to and whether it is too high for his population.

** Evidence provided to support measure

** Beer's list data driving evidence. Grade assigned to evidence by Amer Psych Assoc Practice Guidelines: Moderate

1b. Performance Gap

Comments:

** A performance gap was demonstrated in terms of both overall performance as well as disparities (dual eligible higher rate, age, and sex).

** The performance gap is not high but (6-9%) but this reviewer is concerned that Medicare and Medicaid have the highest rates of antipsychotic ordering for this population suggest that there are disparities in use of antipsychotic by race and gender. Inappropriate use of antipsychotics in inpatient settings to quell a variety of behaviors should be minimized regardless of whether it is continued on an outpatient basis after discharge--though the latter is of great concern.

** Performance gap discussed regarding patient payment method as well as gender. Gender may be related to symptoms being treated by medication.

** Yes, disparities by gender, race and insurance type (Medicare/Medicaid)

** well established

** Developers provide information on a study of three tests sites. These sites evidenced variation use of antipsychotics (19.4%-5.5%) for the targeted population. Overall, the sites included 137,817 hospital encounters. Additionally, a couple of other studies were cited that found variation by age group, and also on continuation of antipsychotics post discharge. Disparities also appear to exist between insurance type (government sponsored vs. private), with Medicaid and Medicare beneficiaries being prescribed more antipsychotics than those with private insurance.
** Performance results were at the facility level from 2 health systems providing data from 10 hospitals and 1critical access hospital. Three test sites, Texas, North Carolina and Pennsylvania, provided the data from 2 EHRs. However, the EHRs were different as were the bed size of the hospitals. Patient disparities were taken into consideration but provider disparity and geography were not. The data does show a gap in care which this measure would address.
** Yes, data demonstrates that there is variability in prescriber habits and that there is likely room for improvement that this quality measure could address. However, evidence didn't seem to point to a "target" goal for prescribing antipsychotics so it is difficult to predict what the right percent of prescription rate is.
** Current performance was provided.
** Lower the score the better; the min of 5.5% is being driven by findings from one very low volume institution compared to the other two - this seems to be driving the range breakdown. It seems there should be a weighted average of the test site scores to then develop a range. It appears the percentile separation points may only be partial percentage points. What is the optimal percentage? I understand the goal is to decrease inappropriate prescribing of antipsychotics during inpt to then reduce outpt use as well - but, what is considered the optimal % of use? I could not find that in the background information.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability**;** Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously.
Feasibility Testing	The submission contains a feasibility assessment that addresses data element feasibility and follow-up with measure developer indicates that the measure logic is feasible based on assessment by EHR vendors

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluation of Reliability and Validity (and composite construction, if applicable):

Link A

Additional Information regarding Scientific Acceptability Evaluation (*if needed*): N/A

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The NQF Staff is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The NQF Staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Scientific Acceptability

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.**

Measure Number: 3315

Measure Title: Use of Antipsychotics in Older Adults in the Inpatient Hospital Setting

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?
   NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

   TIPS: Consider the following: 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?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒Yes (go to Question #6)

   ☐No (please explain below then go to Question #8)

   Split-half correlation (test-retest) was used to assess reliability of the performance measure scores.

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

   Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☒High (go to Question #8)

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☒Yes (go to Question #9)

   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☒Yes (go to Question #10)

   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

    ☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

    ☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☒High (NOTE: Can be HIGH only if score-level testing has been conducted)

☐Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒Yes (go to Question #2)

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable***

threats should result in an overall INSUFFICENT rating for validity, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

   ☐Yes (please explain below then go to Question #3)

   ☒No (go to Question #3)

   ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?

   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

   ☐Yes (please explain below then go to Question #4)

   ☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

   ☐Yes (please explain below then go to Question #5)

   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

   ☐Yes (please explain below then go to Question #6)

   ☒No (go to Question #6)

   ☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?

   ☐Yes (please explain below then go to Question #7)

   ☒No (go to Question #7)

ASSESSMENT OF MEASURE TESTING

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

☐No (please explain below then go to Question #8)

Data element validity testing evaluated whether the measure specification correctly identifies all the data elements required to calculate the measure score. This method quantifies the percent agreement, Kappa statistic, sensitivity, specificity, negative predictive value and positive predictive value between the electronically extracted EHR data and the manually abstracted data (which use the entire record, including free text notes fields).

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

   ☐Yes (go to Question #9)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☒Yes (go to Question #11)

    ☐No (please explain below and go to Question #13)

    Data Element validity testing evaluated whether the measure specification correctly identifies all the data elements required to calculate the measure score. Data element validity was tested by selection of a random set of patient encounters from the EHR extract and comparing data to those that were manually abstracted for the same encounters.

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☒Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

    Validity of performance score was evaluated by surveying test sites and Expert Work Group.

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #14)

    ☒Moderate (go to Question #14)

    ☐Low (please explain below then go to Question #13)

    ☐Insufficient

13. Was other validity testing reported?

    ☐Yes (go to Question #14)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

    TIPS: Prior validity studies of the same data elements may be submitted

    ☒Yes (go to Question #15)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☒Yes (go to Question #16)

    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the

score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

The Kappa values calculated through data element validity testing suggest that data in the EHR accurately reflect patient care. In addition, face validity appears to be high as well. Six out of eight respondents reported that hospitals would score well on the measure if they consistently documented "threat of harm" and denominator exclusions.

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

Comments:

** I share the concern of the TEP that the measure assesses medication ordered rather than actually administered. The specification for how to assess documentation of threatening harm to self or others is unclear to me. Distinguishing ICU from non-ICU care makes sense.
** This eMeasure is well specified and measure logic has been demonstrated reliably.
** Atypical versus typical antipsychotics not addressed.
** Ok
** good
** All data elements are clearly defined including specific codes. The measure could easily be consistently implemented because it can be extracted from codes already in use.
** Satisfied, high reliability.
** Wonder about exclusionary criteria and whether recent suicidal, self-harm, or violent behaviors towards others should be exclusionary criteria.
** Clearly defined specifications.
** Num and Den defined. Not familiar with use of OID for drugs, would RxNorm be a better coding system to identify drugs?

2a2. Reliability – Testing

Comments:

** Split half correlation was used to assess the reliability of the performance measure scores due to sample size and smaller number of test sites. Reliability coefficient was 0.981 with 95% confidence interval, which is acceptable.
** No concerns about reliability. No need to review or discuss in Committee.
** No concern.
** no
** It wasn't clear if any of the hospitals in which the measure was tested had inpatient psychiatric units.
** No concerns. Reliability coefficients were high for the measure.
** No concerns
** If antipsychotics are useful for agitated patients (at risk of harm to self or others) then the list of exclusionary criteria doesn't seem to be sufficient. Many will be captured in numerator that perhaps are rightfully given antipsychotics.
** The logic model is poor as it leaps from reducing fewer patients prescribed antipsychotics to reducing death without describing that pathway more clearly.
** There should be a stronger decision tree indicating which patients antipsychotics are appropriate for - data suggests that for those for whom behavioral interventions haven't been sufficient, so perhaps there needs to be more documentation or decision making such as diagnosis of non-suicidal self injury, recent suicide attempt as also being exclusionary criteria.
** Tested with hospitals that can have data extracted electronically which improves reliability. Score was very high at 0.981. What happens for institutions where data is not readily available in electronic format? Unclear if reliability would be the same; would think so due to objective data driving decision.

2b1. Validity –Testing

2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:
** Face validity was assessed as well as data element validity (98% agreement between EHR and manual abstraction). Comments also evaluated and generally were favorable.
** No concerns about validity. No need to discuss and review in Committee.
** No concern.
** moderately good
** psychiatric diagnoses may be under-recorded in the medical record.
** No concerns. Validity testing results suggest the measure has high validity.
** Satisfied, moderate validity. Only concerns were mentioned above with respect to size, location and differing EHRs.
** Face validity done (8/8 agrees that measure components were appropriate). Data validity also done with 98.1% agreement for all denom and num metrics.
2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment:
Comments:
** Exclusions made sense based on evidence and do not dramatically change results.
** Exclusions listed are not very specific. Nothing mentioned in exclusions about treating behavioral disturbances, delirium, dementia.
** The measure excludes patients on inpatient psychiatric units. I don't know if the developers assumed that if patients with dementia get behaviorally out of control in dangerous ways, they would get transferred to pysch units and hence excluded. the reality is that there still could be an acute need on medical surgical units to use antipsychotics and access to psych units is quite delayed because they're full.
** exclusions align with evidence and there doesn't appear to be any groups excluded from measure inappropriately.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent
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.

- There are no fees or licensing requirements to use this measure – it is in the public domain.
- All data elements used to compute the measure are in defined fields in electronic health records and are generated or collected by and used by healthcare personnel during the provision of care.
- Feasibility Scorecard assesses data elements across three sites and two EHR systems. Overall scores:
    - Data availability is 87 percent
    - Data accuracy is 73 percent
    - Data standards is 67 percent
    - Workflow is 47 percent

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

Comments:

** Generally feasible although not clear that specification of harm to self or others is routinely generated (but could be). Data comes from EHR fields.

** Data demonstrate feasibility in terms of availability of data elements and to a bit less to data accuracy. However, this reviewer has a concern about efforts that need to be made to include use of this measure at the appropriate time in the workflow of different inpatient settings. Studies of other measures have shown that lack of attention to workflow can seriously inhibit use of a measure and data related to use of this measure suggest that is happening.

** Exclusions need to be more clearly defined to help determine feasibility.

** Defined fields in HER

** It's feasible

** All data elements are collected in routine care. All data elements are captured electronically.

** All of the data elements used to compute the measure are in defined fields in the EHR and are generated or collected by and used by healthcare personnel during provision of care.

** Feasibility seems high.

** Should be very feasible.

** all data can be extracted electronically (where available in this format), thus feasibility is improved.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

The measure has been submitted through the Measures Under Consideration process for the CMS Hospital Inpatient Quality Reporting Program. And Medicare and Medicaid Programs; Electronic Health Record Incentive Program – Stage 3.

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

N/A – this measure has not been implemented.

Additional Feedback (including NQF member support/non-support of the measure):

N/A

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.
4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.
Improvement results [Impact/trends over time/improvement]
The developer provides rational for how the performance results could be used to support high quality care:

- Adoption of this performance measure has the potential to improve the quality of care for hospitalized older adults in the area of patient safety.
- Encourage thoughtful prescribing of antipsychotics for hospitalized patients result in fewer prescriptions after discharge reducing morbidity and mortality associated with long-term use of these medications.

4b2. Benefits vs. harms. 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).
Unexpected findings (positive or negative) during implementation

- Thoughtful prescribing of antipsychotics
- Encourage delirium assessment and monitoring tools
- Use of non-pharmacologic interventions

Potential harms
Potential increased use of alternative harmful medications such as benzodiazepines for delirium or BPSD.
Additional Feedback:

N/A

***Questions for the Committee**:*

- *How 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?*

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE:

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency
Comments:
** Being considered by CMS for quality reporting.
** Measure is not currently being reported for accountability. Since it is not being used currently, there is no feedback. The measure has been used in retrospective studies in real-world settings.
** Measure is being considered for use for the CMS Hospital Inpatient Quality Reporting Program.

** I am not sure how systems will be able to interpret results to make comparisons to know what target or reduction to aim for.
** Measure submitted for CMS hospital inpatient quality reporting program. Unclear on actual use when target % for appropriate use is not defined - how far off are we with average of 20% use?

4b1. Usability – Improvement

Comments:

** There is concern of potential unintended consequences of driving up use of benzodiazepine medication or increase in physical restraints.
** The benefits of use of this measure should increase assessment and monitoring of patients before use of any antipsychotic and should encourage use of non-pharmacologic interventions prior of consideration of use of an antipsychotic.
** Unsure that this data provides any direct link to inappropriate use of antipsychotic medications in elderly population. Possible that effective use of this medications will be curtailed by use of this measure.
** Potential to improve.
** I think that some developer might consider harmonizing this with outpatient, post acute care, and nursing home use.
** The potential harms are that patients who need or could benefit from antipsychotic medications may not receive them, patients who were on antipsychotic medications when admitted may have their medications inappropriately discontinued, patients behavioral symptoms may be managed by restraint or using other inappropriate medications.
** Use of the measure could result in improved quality of care for hospitalized non-psychiatric patients, resulting in reductions in adverse events. The measure may increase the use of behavioral interventions and reduce the unnecessary use of antipsychotic medication.
** This measure would help shift the focus from the use of medications, specifically antipsychotics, to determining alternative causes of behavior and adjusting treatment accordingly. The benefits outweigh any harm.
** Seems a lot of individuals are started on antipsychotics in the hospital due to increased delirium or agitation and that is what this measure is trying to avoid. However, seems that this is really more about a continuity of care issue and that it is the outpatient provider's responsibility to ensure weaning/changes of the meds as the patient's health improves.
** Very useable and can be obtained from EHRs
** Potential harm if hospitals get payment withhold: Patients who need the drug could potentially be prohibited from getting drug (unintended negative consequence).

Criterion 5: Related and Competing Measures

Related or competing measures

2111 : Antipsychotic Use in Persons with Dementia
2993 : Potentially Harmful Drug-Disease Interactions in the Elderly

Harmonization

Developer states that the measures are harmonized to the extent possible. The submitted measure is the only inappropriate use measure that assesses use of antipsychotic medications in the inpatient hospital setting.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January 10, 2018

- The American Academy of Neurology in general supports the measure, but notes the following potential concerns. Was an exclusion/exception considered for patients that pose harm to themselves or others? There is concern the population impacted will be small given published rates of between 9-6% for use of antipsychotic medications in inpatient settings. Defined outcomes are general and will be difficult to link to the measure as opposed to other factors that might effect post-hospital morbidity and mortality. Finally, reducing unnecessary continuation of antipsychotics following discharge would be a more tangible outcome.
- No NQF Members have submitted support/non-support choices as of this date.

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

AP_Evidence-636451281689967368.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Not applicable

Measure Title: Use of Antipsychotics in Older Adults in the Inpatient Hospital Setting

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Not applicable

Date of Submission: 11/1/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- 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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

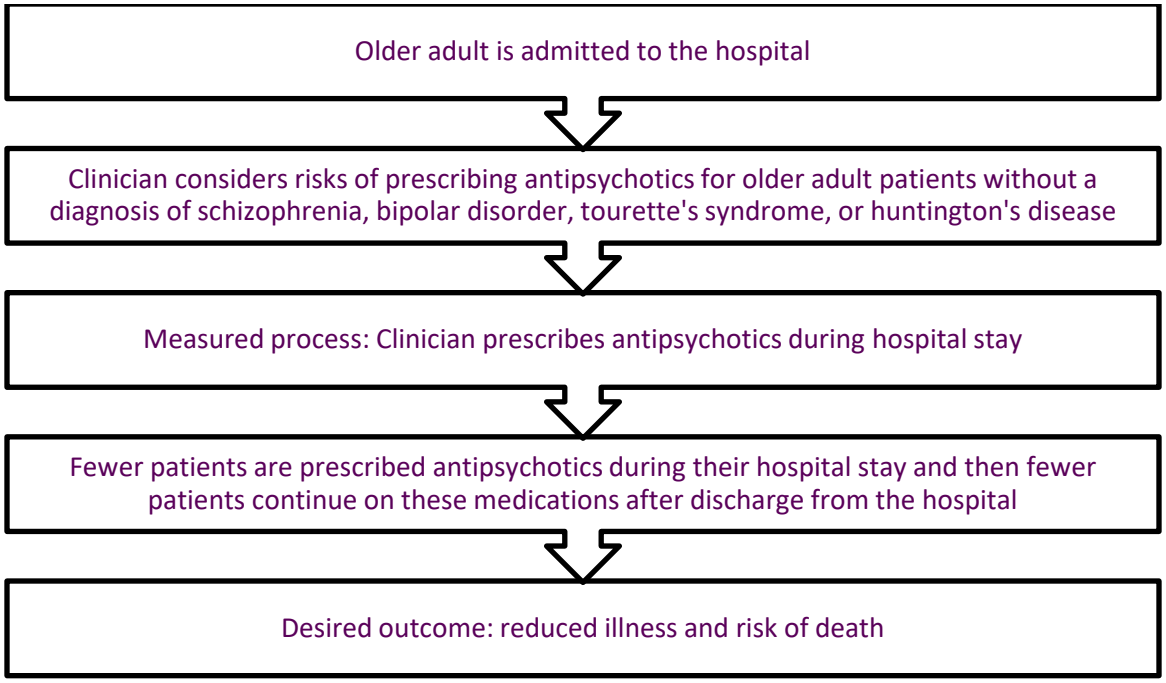
☒ Process: Prescribing of potentially inappropriate medications for older adults

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.



1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

Not applicable.

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☒ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Table 1. American Geriatrics Society Guideline

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: American Geriatrics Society 2015 Updated Beers Criteria for Potentially Inappropriate Medication Use in Older Adults.** **Author: American Geriatrics Society 2015 Beers Criteria Update Expert Panel.** **Date: 2015.** **Citation: American Geriatrics Society 2015 Beers Criteria Update Expert Panel. 2015. American Geriatrics Society 2015 Updated Beers Criteria for Potentially Inappropriate Medication Use in Older Adults. Journal of the American Geriatrics Society, 63(11): 2227-2246.** **URL: http://geriatricscareonline.org/ProductAbstract/american-geriatrics-society-updated-beers-criteria-for-potentially-inappropriate-medication-use-in-older-adults/CL001**

Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**Table 2** 2015 American Geriatrics Society Beers Criteria for Potentially Inappropriate Medication Use in Older Adults **Organ System, Therapeutic Category, Drugs:** Antipsychotics, first- (conventional) and second- (atypical) generation **Rationale:** Increased risk of cerebrovascular accident (stroke) and greater rate of cognitive decline and mortality in persons with dementia Avoid antipsychotics for behavioral problems of dementia or delirium unless nonpharmacological options (e.g., behavioral interventions) have failed or are not possible and the older adult is threatening substantial harm to self or others **Recommendation:** Avoid, except for schizophrenia, bipolar disorder, or short-term use as antiemetic during chemotherapy **Table 3** 2015 American Geriatrics Society Beers Criteria for Potentially Inappropriate Medication Use in Older Adults Due to Drug-Disease or Drug-Syndrome Interactions That May Exacerbate the Disease or Syndrome **Disease or Syndrome:** Delirium **Drug(s):** Antipsychotics **Rationale:** Avoid in older adults with or at high risk of delirium because of the potential of inducing or worsening delirium Avoid antipsychotics for behavioral problems of dementia or delirium unless nonpharmacological options (e.g., behavioral interventions) have failed or are not possible and the older adult is threatening substantial harm to self or others Antipsychotics are associated with greater risk of cerebrovascular accident (stroke) and mortality in persons with dementia **Recommendation:** Avoid **Table 3** 2015 American Geriatrics Society Beers Criteria for Potentially Inappropriate Medication Use in Older Adults Due to Drug-Disease or Drug-Syndrome Interactions That May Exacerbate the Disease or Syndrome **Disease or Syndrome:** Dementia or cognitive impairment **Drug(s):** Antipsychotics, chronic and as-needed use **Rationale:** Avoid because of adverse CNS effects Avoid antipsychotics for behavioral problems of dementia or delirium unless nonpharmacological options (e.g., behavioral interventions) have failed or are not possible and the older adult is threatening substantial harm to self or others. Antipsychotics are associated with greater risk of cerebrovascular accident (stroke) and mortality in persons with dementia **Recommendation:** Avoid

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**Moderate**: Evidence is sufficient to determine risks of adverse outcomes, but the number, quality, size, or consistency of included studies; generalizability to routine practice; or indirect nature of the evidence on health outcomes (≥1 higher-quality trial with >100 participants; ≥2 higher-quality trials with some inconsistency; ≥2 consistent, lower-quality trials; or multiple, consistent observational studies with no significant methodological flaws showing at least moderate effects) limits the strength of the evidence
Provide all other grades and definitions from the evidence grading system	High: Evidence includes consistent results from well designed, well-conducted studies in representative populations that directly assess effects on health outcomes (≥2 consistent, higher-quality randomized controlled trials or multiple, consistent observational studies with no significant methodological flaws showing large effects) Low: Evidence is insufficient to assess harms or risks in health outcomes because of limited number or power of studies, large and unexplained inconsistency between higher-quality studies, important flaws in study design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes
Grade assigned to the **recommendation** with definition of the grade	**Strong**: Benefits clearly outweigh harms, adverse events, and risks, or harms, adverse events, and risks clearly outweigh benefits
Provide all other grades and definitions from the recommendation grading system	Weak: Benefits may not outweigh harms, adverse events, and risks Insufficient: Evidence inadequate to determine net harms, adverse events, and risks

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The Beers Criteria were first published in 1991. Since that time the criteria have been regularly updated based off of the existing criteria and any new evidence published since the last update. The American Geriatrics Society forms an expert panel to update the Beers Criteria every few years. The panel works from the previous evidence review and then reviews any new evidence published since that last review to update the recommendations in the Beers Criteria. The 2015 review by the AGS 2015 Beers Criteria Update Expert Panel included review of 60 systematic reviews and meta analyses, 49 randomized control trials (RCTs) and 233 observational studies and other types of publications. Overall, the quality of the evidence is good. In addition to conducting a systematic review of the evidence, the AGS 2015 Beers Criteria Update Expert Panel also used technical experts and a public comment period for additional validity. **Table 2, antipsychotics:** Evidence for the recommendation to avoid antipsychotics in older adults was rated as moderate quality. It includes 2 randomized control studies, 3 systematic reviews, 2 cohort studies and 1 observational study. **Table 3, delirium:** Evidence for the recommendation to avoid certain medications (including antipsychotics) for individuals with delirium was rated as moderate quality. It includes 2 systematic reviews, 1 randomized controlled study, 8 cohort studies, 1 observational study and 1 clinical review. **Table 3, dementia or cognitive impairment**: Evidence for the recommendation to avoid certain medications (including antipsychotics) for individuals with dementia was rated as moderate quality. It includes 3 systematic reviews and 2 randomized control studies in addition to 4 cohort studies.

Estimates of benefit and consistency across studies	Recommendations in the Beers criteria are based on studies that explain the rationale for why a medication group is potentially harmful for older adults (Table 2) or for older adults with a certain condition (Table 3). Below is a summary of the number and types of studies supporting the relevant recommendations regarding antipsychotics. Studies consistently found an increased risk of adverse events associated with antipsychotic use. Summaries of each study can be found on the American Geriatrics Society's website: http://www.americangeriatrics.org/. **Table 2, Antipsychotics** Studies that support the recommendation: 2015 Criteria: Hwang 2014 – retrospective cohort Langballe 2014 – retrospective cohort From previous criteria: Dore 2009 – observational Maher 2011 – systematic review, meta-analysis Schneider 2005 – systematic review, meta-analysis Schneider 2006a – systematic review, meta-analysis Schneider 2006b – randomized control trial Vigen 2011 – randomized control trial Recommendation: Avoid, except for schizophrenia, bipolar disorder, or short-term use as antiemetic during chemotherapy **Table 3, Delirium** Studies that support the recommendation: From 2015 Criteria: Aparasu 2012 – retrospective cohort Chavant 2011 – retrospective cohort Citrome 2013 – retrospective cohort Hampton 2014 – retrospective cohort Han 2004 – randomized control trial Rigler 2013 – retrospective cohort From previous criteria: Clegg 2011 – systematic review Gaudreau 2005 – prospective cohort Laurila 2008 – observational study Marcantonio 1994a – prospective cohort Moore 1999 – clinical review Ozbolt 2008 – systematic review Rudolph 2008 – retrospective and prospective cohorts Recommendation: Avoid in older adults with or at high risk of delirium because of potential of inducing or worsening delirium Avoid antipsychotics for behavioral problems of dementia and/or delirium unless nonpharmacological options (e.g., behavioral interventions) have failed or are not possible and the older adult is threatening substantial harm to self or others. Antipsychotics are associated with increased risk of cerebrovascular accident (stroke) and mortality in persons with dementia **Table 3, Dementia or cognitive impairment** Studies that support the recommendation: 2015 Criteria: Chavant 2011 – retrospective cohort

	Kalicsh Ellet 2014 – retrospective cohort From previous criteria: Rudolph 2008 – retrospective and prospective cohorts Schneider 2005 – systematic review, meta-analysis Schneider 2006a – systematic review, meta-analysis Schneider 2006b – randomized control trial Seitz 2011 – systematic review, meta-analysis Vigen 2011 – randomized control trial Wright 2009 – prospective longitudinal cohort Recommendation: Avoid due to adverse CNS effects Avoid antipsychotics for behavioral problems of dementia and/or delirium unless nonpharmacological options (e.g., behavioral interventions) have failed or are not possible and the older adult is threatening substantial harm to self or others. Antipsychotics are associated with increased risk of cerebrovascular accident (stroke) and mortality in persons with dementia
What harms were identified?	As part of their review of the evidence, the AGS 2015 Beers Criteria Update Expert Panel identified subgroups of patients who should be exempt from the criteria and for whom listed medications may be appropriate. In addition, a patient could have a condition or comorbidity that would merit the use of a medication on the list, even if the comorbidity is not specifically listed in the criteria. The criteria are designed to assist providers in the prescribing of potentially harmful medications, and should not be taken as strict criteria to avoid use in all patients without weighing the harms and benefits for individual cases.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Relevant studies have been published since the publication of the guideline, but they do not change these conclusions. Relevant studies include, but are not limited to, the following: Herzig SJ, Rothberg MB, Guess JR, et al. 2016. "Antipsychotic Use in Hospitalized Adults: Rates, Indications, and Predictors." J Am Geriatr Soc 64(2): 299-305. doi: 10.1111/jgs.13943 Marshall J, Herzig SJ, Howell MD, et al. 2016. "Antipsychotic utilization in the intensive care unit and in transitions of care." J Crit Care 33: 119-124. doi: 10.1016/j.jcrc.2015.12.017 Neufeld KJ, Yue J, Robinson TN, et al. 2016. "Antipsychotic Medication for Prevention and Treatment of Delirium in Hospitalized Adults: A Systematic Review and Meta-Analysis." J Am Geriatr Soc 64(4):705-714. doi: 10.1111/jgs.14076

Table 2. American Psychiatric Association Guideline

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Title: The American Psychiatric Association Practice Guideline on the Use of Antipsychotics to Treat Agitation or Psychosis in Patients with Dementia** **Author: Reus VI, Fochtmann LJ, Eyler AE, Hilty DM, Horvitz-Lennon M, Jibson MD, Lopez OL, Mahoney J, Pasic J, Tan ZS, Wills CD, Rhoads R, Yager J.** **Date: 2016** **Citation, including page number: Reus VI, Fochtmann LJ, Eyler AE, et al. 2016. "The American Psychiatric Association Practice Guideline on the Use of Antipsychotics to Treat Agitation or Psychosis in Patients with Dementia." *Am J Psychiatry* 173(5):543-6. doi: 10.1176/appi.ajp.2015.173501.** **URL: http://ajp.psychiatryonline.org/doi/pdf/10.1176/appi.ajp.2015.173501**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	"Statement 5. APA recommends that nonemergency antipsychotic medication should only be used for the treatment of agitation or psychosis in patients with dementia when symptoms are severe, are dangerous, and/or cause significant distress to the patient. (1B)"
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**"Moderate** (denoted by the letter B) = Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate."
Provide all other grades and definitions from the evidence grading system	"High (denoted by the letter A) = High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect." "Low (denoted by the letter C) = Low confidence that the evidence reflects the true effect. Further research is likely to change our confidence in the estimate of effect and is likely to change the estimate."
Grade assigned to the **recommendation** with definition of the grade	**"Recommendation"** (denoted by the numeral 1 after the guideline statement) indicates confidence that the benefits of the intervention clearly outweigh harms.
Provide all other grades and definitions from the recommendation grading system	"Suggestion" (denoted by the numeral 2 after the guideline statement) indicates uncertainty (i.e., the balance of benefits and harms is difficult to judge or either the benefits or the harms are unclear).
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Overall, 45 randomized controlled trials and 52 observational studies were included in the guideline.

Estimates of benefit and consistency across studies	"Statements 5, 8, 10, 14, and 15 are based on moderate-strength evidence in individuals with dementia that the benefits of antipsychotic medication are small. In addition, consistent evidence, predominantly from large observational studies, indicates that antipsychotic medications are associated with clinically significant adverse effects, including mortality, among individuals with dementia. The overall strength of evidence for these statements is graded as moderate on the basis of this balance of benefits and harms data and the fact that there were no studies that directly addressed all of the specific elements of each recommendation."
What harms were identified?	The Guideline Writing Group acknowledged that there are some situations where antipsychotic use for patients with dementia may be appropriate: "Expert consensus suggests that use of an antipsychotic medication in individuals with dementia can be appropriate, particularly in individuals with dangerous agitation or psychosis (see "Expert Opinion Survey Data: Results" in Appendix B), and can minimize the risk of violence, reduce patient distress, improve the patient's quality of life, and reduce caregiver burden. However, in clinical trials, the benefits of antipsychotic medications are at best small (Corbett et al. 2014; Kales et al. 2015; see "Review of Supporting Research Evidence" in Appendix A) whether assessed through placebo controlled trials, head-to-head comparison trials, or discontinuation trials."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Relevant studies have been published since the publication of the guideline, but they do not change these conclusions. Relevant studies include, but are not limited to, the following: Herzig SJ, Rothberg MB, Guess JR, et al. 2016. "Antipsychotic Use in Hospitalized Adults: Rates, Indications, and Predictors." J Am Geriatr Soc 64(2): 299-305. doi: 10.1111/jgs.13943 Marshall J, Herzig SJ, Howell MD, et al. 2016. "Antipsychotic utilization in the intensive care unit and in transitions of care." J Crit Care 33: 119-124. doi: 10.1016/j.jcrc.2015.12.017 Neufeld KJ, Yue J, Robinson TN, et al. 2016. "Antipsychotic Medication for Prevention and Treatment of Delirium in Hospitalized Adults: A Systematic Review and Meta-Analysis." J Am Geriatr Soc 64(4):705-714. doi: 10.1111/jgs.14076

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Clinical guidelines recommend against using antipsychotics as a standard first line of treatment for patients experiencing delirium or behavioral and psychological symptoms of dementia unless they present a threat to themselves or their caregivers (AGS 2015a, AGS 2015b, NICE 2016, Reus 2016). Antipsychotics are often used off-label as a method of treating patients in an acute confusional state despite conflicting evidence regarding the effectiveness of antipsychotics in treating these disorders (Neufeld 2016, Barr 2013, Campbell 2009). The benefits of this measure lie in the potential to reduce inappropriate use of antipsychotics in inpatient hospital settings and the unnecessary continuation of the intervention post-discharge, resulting in improved patient outcomes (reduced morbidity and mortality) for older adults. Measuring the use of antipsychotics among hospitalized older adult patients could help shift the focus to determining alternative causes of this behavior (such as medication interactions, medication side effects, etc.) and adjusting treatment accordingly.

American Geriatrics Society 2015 Beers Criteria Update Expert Panel. (2015a). American Geriatrics Society 2015 Updated Beers Criteria for Potentially Inappropriate Medication Use in Older Adults. J Am Geriatr Soc, 63(11), 2227-

AGS Expert Panel on Postoperative Delirium in Older Adults. "American Geriatrics Society abstracted clinical practice guideline for postoperative delirium in older adults." J Am Geriatr Soc, 63(1), 2015b, pp 142-50. doi: 10.1111/jgs.13281.

Barr, J., Fraser, G.L., Puntillo, K., et al. (2013). Clinical practice guidelines for the management of pain, agitation, and delirium in adult patients in the intensive care unit. Crit Care Med, 41(1), 263-306.

Campbell, N., Boustani, M.A., Ayub, A., et al. (2009). Pharmacological management of delirium in hospitalized adults--a systematic evidence review. J Gen Intern Med, 24(7), 848-53.

Neufeld, K.J., Yue, J., Robinson, T.N., et al. (2016). Antipsychotic Medication for Prevention and Treatment of Delirium in Hospitalized Adults: A Systematic Review and Meta-Analysis. J Am Geriatr Soc, 64(4), 705-714.

NICE (National Institute for Health and Clinical Excellence) Dementia: Supporting people with dementia and their carers in health and social care. 2016 (Issued November 2006, Modified September 2016).

Reus, V.I., Fochtmann, L.J., Eyler, A.E., et al. (2016). The American Psychiatric Association Practice Guideline on the Use of Antipsychotics to Treat Agitation or Psychosis in Patients with Dementia. Am J Psychiatry, 173(5), 543-546.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

The measure was tested in two health systems, which provided data from 10 hospitals, and one critical access hospital. These systems, Test Site 1 (in Texas), Test Site 2 (in North Carolina), and Test Site 3 (in Pennsylvania) varied in terms of their EHR product and the size of their hospital systems. Test Sites 1 and 2 used different installations of the Cerner EHR product. Test Site 3 used a Meditech EHR product. With respect to size, Test Site 1 had the most beds (n=3,320) and Test Site 3, a critical access hospital, had the least number of beds (n=25). Test Site 1 provided data for nine hospitals in its system. Test Site 2 and Test Site 3 provided data for one hospital each. Across the three test sites, we received data on 137,817 hospital encounters. Test Site 1 contributed the most encounters (n= 99,528) followed by Test Site 2 (n=37,560) and Test Site 3 (n=729). A detailed breakdown of the characteristics of the measured facilities and the patient populations can be found in sections 1.5 and 1.6 of the attached Measure Testing form.

The measure performance, including the denominator, numerator, and the measure rate by hospital, is presented below.

Test site 1:

- Dates of data: October 1, 2013–September 30, 2015
- Denominator: 99,528
- Denominator after exclusions: 92,943
- Numerator: 16,229
- Numerator exclusions: 153
- Measure rate: 17.3%

Test site 2:

- Dates of data: October 1, 2013–September 30, 2015
- Denominator: 37,560
- Denominator after exclusions: 35,385
- Numerator: 6,984
- Numerator exclusions: 112
- Measure rate: 19.4%

Test site 3:

- Dates of data: October 1, 2014–September 30, 2015
- Denominator: 729
- Denominator after exclusions: 727
- Numerator: 40
- Numerator exclusions: 0
- Measure rate: 5.5%

Overall: (summary statistics based on the October 1, 2014–September 30, 2015, data from Test Sites 1 & 2 and October 1, 2014–September 30, 2015, data from Test Site 3)

- Mean: 15.3%
- Std. Deviation: 5.2%
- Coefficient of variation: 0.342
- Min: 5.5%
- Max: 22.8%
- Interquartile Range: 5.9%
- 10th Percentile: 8.4%

- 25th Percentile: 12.8%

- 50th Percentile: 15.9%

- 75th Percentile: 18.7%

- 90th Percentile: 20.4%

Overall: (summary statistics based on the October 1, 2014–September 30, 2015, data from Test Sites 1 & 2)

- Mean: 16.3%

- Std. Deviation: 4.3%

- Coefficient of variation: 0.265

- Min: 8.4%

- Max: 22.8%

- Interquartile Range: 5.1%

- 10th Percentile: 11.4%

- 25th Percentile: 14.0%

- 50th Percentile: 16.9%

- 75th Percentile: 19.1%

- 90th Percentile: 20.7%

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.

Research on the use of antipsychotics during an inpatient hospital visit is limited, but there are indications of a quality gap. Recent studies have estimated the prevalence of potentially inappropriate antipsychotic use in the inpatient setting.

A retrospective cohort study of roughly 18,000 adult non-psychiatric hospital admissions over a year found antipsychotic exposure in 9 percent of visits. More than half of these were patients who may have been initiated on antipsychotics during those visits. 26 percent of the patients who were initiated on an antipsychotic were then discharged on an antipsychotic. The most common reasons documented for initiating antipsychotics were delirium or probable delirium (Herzig 2016a).

A retrospective cohort study of 2,700,000 adult non-psychiatric hospital admissions over a year found antipsychotic exposure in 6 percent of visits. This rate varied by age, with 4.6 percent of patients age 18–65, 5.2 percent of patients age 65–74, and 8.8 percent of patients age 75 and older being exposed to antipsychotics during their inpatient stay. This study also found that 29 percent of admissions with delirium and 27 percent of admissions with dementia received antipsychotics. This study concluded that there was variation in antipsychotic use between hospitals, which should be explored further (Herzig 2016b). A retrospective cohort study of approximately 39,000 ICU admissions over the course of 7 years found that 8 percent of patients were newly initiated on antipsychotics during the ICU visit and 21 percent of these patients were continued on antipsychotics after discharge (Marshall 2016).

Herzig, S.J., M.B. Rothberg, J.R. Guess, et al. “Antipsychotic Use in Hospitalized Adults: Rates, Indications, and Predictors” J Am Geriatr Soc, 64(2), 2016a, pp 299-305.

Herzig, S.J., M.B. Rothberg, J.R. Guess, et al. “Antipsychotic medication utilization in nonpsychiatric hospitalizations.” J Hosp Med, 11(8), 2016b, pp 543-549.

Marshall, J., S.J. Herzig, M.D. Howell, et al. “Antipsychotic utilization in the intensive care unit and in transitions of care.” J Crit Care, 33, 2016, pp 119-24.

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 maintenance*

of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.

Data collected during measure testing on older adult patients (65 years and older) found that patients with Medicare and Medicaid coverage had the highest rate of antipsychotic ordering (22.0 percent and 27.9 percent, respectively). Patients with private insurance had the lowest rates at 13.4 percent. Measure testing found statistically significant differences in antipsychotic orders for males compared to females (24.0 and 19.7 percent, respectively). Difference in the rate of antipsychotic ordering by race is significant as well. Across racial groups (black, white, and other), the rate of antipsychotic ordering ranged from 20.9 to 24.4. Hispanic and non-Hispanic patients had similar performance rates, 20.6 and 22.0 respectively. Although the difference is small between the two ethnicity groups and likely not clinically significant, it is statistically significant (p=.022), which is likely due to the large sample size.

Data collected during measure testing on all adult patients found that patients with Medicare and Medicaid coverage had the highest rate of antipsychotic ordering (21.0 percent and 20.1 percent, respectively). Patients with private insurance had the lowest rates at 11.3 percent. Measure testing found statistically significant differences in antipsychotic orders for patients 65 and older compared to patients age 18–64 (21.6 versus 14.8, respectively) and significant differences in males compared to females (19.8 and 16.1 percent, respectively). There was little difference in the rate of antipsychotic ordering by race or ethnicity. Across racial groups (black, white, and other), the rate of antipsychotic ordering ranged from 17.6 to 18.4. Hispanic and non-Hispanic patients had similar performance rates, 16.6 and 18.1 respectively. Although these differences are small and likely not clinically significant, they are statistically significant (p<.001). This is likely due to the large sample size. (Barrett 2017).

Barrett, K,. F. Xing, K. Sobel, and B. Rehm. "Hospital Inpatient and Outpatient Process and Structural Measure Development and Maintenance Project: Beta Testing Report on the Use of Antipsychotics in Adults in the Inpatient Hospital Setting Electronic Clinical Quality Measure." Washington, DC: Mathematica Policy Research, July 2017.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

The research on disparities in the use of antipsychotics is limited. According to one researcher, factors such as insurance status and race have been associated with the use of antipsychotics in hospitalizations. Patients with Medicare, Medicaid, or self-pay primary insurance are more likely to receive antipsychotics than patients with commercial primary insurance (Herzig, 2016a). Herzig also observed that non-white individuals are less likely to receive antipsychotics than white individuals. Further scientific investigation is required to understand the reasons for these disparities (Herzig 2016b).

Herzig, S. J., M. B. Rothberg, et al. (2016a). "Antipsychotic medication utilization in nonpsychiatric hospitalizations." J Hosp Med.

Herzig, S. J., M. B. Rothberg, et al. (2016b). "Antipsychotic Use in Hospitalized Adults: Rates, Indications, and Predictors." J Am Geriatr Soc 64(2): 299-305.

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 sub criteria 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):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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.)*

There is not a measure-specific web page but the specifications are attached in accordance with question S.2a.

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 an eMeasure **Attachment:** AP_eCQMSpecs.zip

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:** AP_ValueSets.xlsx

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Not applicable

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) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Inpatient hospitalizations for patients who received an order for an antipsychotic medication during the inpatient encounter.

S.5. 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, time period for data collection, 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 (S.14).

The time period for data collection is the measurement year (12-month period).

Numerator: Inpatient hospitalizations for patients who received an order for an antipsychotic medication during the inpatient encounter.

Antipsychotic orders are represented with the QDM datatype and value set of Medication, Order: Antipsychotic Medications (OID: 2.16.840.1.113883.3.464.1003.196.12.1255).

Numerator exclusions: Inpatient hospitalizations for patients with documented indication that they are threatening harm to self or others

Threat to self or others is represented with the QDM datatype and value set of Symptom: Threat to themselves or others (OID: 2.16.840.1.113883.3.464.1003.195.12.1020).

To access the value sets for the measure, please visit the Value Set Authority Center, sponsored by the National Library of Medicine, at https://vsac.nlm.nih.gov/. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b.

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Denominator: Non-psychiatric inpatient hospitalizations for patients who are 65 and older.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The time period for data collection is the measurement year (12-month period).

Denominator: Non-psychiatric inpatient hospitalizations for patients who are 65 and older.

Inpatient hospitalizations are represented with the QDM datatype and value set of Encounter, Performed: Encounter Inpatient (OID: 2.16.840.1.113883.3.666.5.3001).

To access the value sets for the measure, please visit the Value Set Authority Center, sponsored by the National Library of Medicine, at https://vsac.nlm.nih.gov/. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Denominator Exclusions: Inpatient hospitalizations for patients with a diagnosis of schizophrenia, Tourette´s syndrome, bipolar disorder, Huntington´s disease during the encounter.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.)*

Denominator Exclusions: Inpatient hospitalizations for patients with a diagnosis of schizophrenia, Tourette´s syndrome, bipolar disorder, Huntington´s disease during the encounter.

Theses exclusions are represented with the QDM datatype of Diagnosis.

- Schizophrenia (OID: 2.16.840.1.113883.3.464.1003.105.12.1104)

- Tourette´s Syndrome (OID: 2.16.840.1.113883.3.464.1003.105.12.1030)

- Bipolar Disorder (OID: 2.16.840.1.113883.3.67.1.101.1.128)

- Huntington´s Disease (OID: 2.16.840.1.113883.3.464.1003.105.12.1032)

To access the value sets for the measure, please visit the Value Set Authority Center, sponsored by the National Library of Medicine, at https://vsac.nlm.nih.gov/. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.)*

Results include a total score and the following strata:

- Stratum 1 - Patients who were admitted or transferred to the ICU during the inpatient encounter

- Stratum 2 - Patients who were not admitted or transferred to the ICU during the inpatient encounter

These strata are identified using the QDM datatype of Encounter, Performed.

ICU Admission or Transfer (OID: 2.16.840.1.113883.17.4077.3.2040)

To access the value sets for the measure, please visit the Value Set Authority Center, sponsored by the National Library of Medicine, at https://vsac.nlm.nih.gov/. A list of value sets for the measure is attached in the Excel workbook provided for question S.2b.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Stratification by risk category/subgroup

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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 = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

Please see the attached HQMF specifications for the complete measure logic. Additionally, a flow diagram of the denominator and numerator logic is attached to the NQF submission form as a supplemental document in response to question A.1, ´AP_LogicFlow_for S.14 response.pdf´.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Not applicable; this measure does not use a sample.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

Not applicable; this measure does not use a survey.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Electronic Health Records

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Hospitals collect EHR data using certified electronic health record technology (CEHRT). The human readable format and XML are contained in the eCQM specifications attached in question S.2a. No additional tools are used for data collection for eMeasures.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. <u>COMPOSITE Performance Measure</u> - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

Not applicable

2. Validity – See attached Measure Testing Submission Form

AP_Testing.docx

2.1 <u>For maintenance of endorsement</u>

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 <u>For maintenance of endorsement</u>

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 <u>For maintenance of endorsement</u>

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** Not applicable
Measure Title: Use of Antipsychotics in Older Adults in the Inpatient Hospital Setting
Date of Submission: 11/1/2017

Type of Measure:

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

Instructions

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

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

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

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

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

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

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

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

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

Not applicable. We did not use an existing data set to test this measure; instead, to test this measure, we partnered with three test sites to extract data from their EHR systems (described in question 1.5). In alignment with the measure's general intent of assessing the use of antipsychotics, we asked hospital staff to submit patient-level data for all patients that qualify for the initial patient population over a one- to two-year period, which includes inpatient admissions for patients 18 years and older (as of the date of the encounter), excluding those with a principal diagnosis of Huntington's, Tourette's, bipolar, or schizophrenia, and where these medications are FDA approved for use. Since the measure is specified for older adults ages 65 and above, all analyses provided in this test report are limited to that age cohort.

1.3. What are the dates of the data used in testing? 10/1/2013-9/30/2015 (Test Site 1 and 2); 10/1/2014 – 9/30/2015 (Test Site 3)

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

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

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*)

Our test data includes data from 11 hospitals across three test sites (two health systems and one critical access hospital). When selecting sites, we ensured representation of at least two different EHR systems across sites, as required by NQF for eCQM testing. We also purposely sought sites whose EHR systems captured the data elements required for the measure calculation and had the ability to create an electronic data extract. By selecting test sites that could provide data for multiple hospitals, we were able to achieve a mix of urban and rural settings and care settings with large and small bed counts. All test sites were non-profit. Test Sites 1 and 2 are teaching hospitals. Test Site 3 is a small, rural safety net hospital. Table 1 lists characteristics of the hospitals participating in field testing.

Table 1. Field testing hospital characteristics

	Hospital	State	Geography	# of beds	EHR product	Inception of current EHR system
Test site 1	All	TX	Urban	3,320	Cerner	2006
	1	TX	Urban	260	Cerner	2006
	2	TX	Urban	877	Cerner	2006
	3	TX	Urban	142	Cerner	2006
	4	TX	Urban	444	Cerner	2006
	5	TX	Urban	255	Cerner	2006
	6	TX	Urban	274	Cerner	2006
	7	TX	Urban	149	Cerner	2006
	8	TX	Urban	568	Cerner	2006
	9	TX	Urban	351	Cerner	2006
Test site 2	10	NC	Urban	874	Cerner	2006
Test site 3	11	PA	Rural	25	Meditech	2010

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*)

Across the three test sites, we received data for 58,507 patient encounters. Across sites, the average age of patients was 76.5 years with a range across sites from a low of 74.6 years to a high of 78.1 years. Distribution by sex was fairly even for Test Sites 1 and 2; at Test Site 3, approximately two-thirds of the patients were female. Across sites, the majority of patients were White and non-Hispanic. At Test Sites 1 and 2, over 90 percent of patients had Medicare. Approximately 60 percent of patients at Test Site 3 had Medicare and 36 percent had private insurance. See Table 2 for a breakdown of these demographic characteristics by test site.

Table 2. Demographic characteristics of the field-testing sample

	Test site 1		Test site 2		Test site 3		Across sites (pooled data)	
Characteristics	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Number of patients	45,097		12,954		456		58,507	
Average age	77.0		74.6		78.1		76.5	
Sex								
Male	18,948	42.0	6,277	48.5	172	37.7	25,397	43.4
Female	26,145	58.0	6,677	51.5	284	62.3	33,106	56.6
Race								
White	28,381	65.7	9,810	76.6	455	99.8	38,446	68.3
Black	6,407	14.8	2,543	19.9	1	0.2	8,951	15.9
Other	8,439	19.5	457	3.6	0	0.0	8,896	15.8
Ethnicity								
Hispanic	4,889	10.8	175	1.4	0	0	5,064	8.7
Non-Hispanic	37,215	82.5	12.404	95.8	436	95.6	50,055	85.6
Other or unknown	2,993	6.7	375	2.8	20	4.4	3,388	5.8
(Primary) Payer								
Medicare	41,415	91.8	11,924	92.1	279	61.2	53,618	91.6
Medicaid	585	1.3	83	.64	9	2.0	677	1.2
Private insurance	2,513	5.6	764	5.9	166	36.4	3,443	5.9
Self-pay or uninsured	283	0.6	63	0.5	0	0.0	346	0.6
Others	301	0.7	120	0.9	2	0.4	423	0.7

SOURCE: Test Sites 1 and 2 data from October 1, 2013 to September 30, 2015. Test Site 3 from October 1, 2014 to September 30, 2015.

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.

- Reliability: We used electronically extracted EHR data from 11 hospitals to examine the reliability of the measure performance rate. Data used were for the time period described in Question 1.3.
- Data element validity: We randomly selected a sample of encounters from each test site's electronic EHR extract and manually abstracted data for those encounters in order to assess the chance-adjusted agreement between the two sources. Manual abstraction was done by trained medical record abstractors. A total of 158 encounters were abstracted across test sites.
- Face validity. We solicited feedback on face validity via interviews and a brief web survey from clinicians, information technology professionals, subject matter experts, and members of the expert workgroup (n=8 respondents).
- Exclusions: We used electronically extracted EHR data from 11 hospitals to examine the impact of the numerator and denominator exclusions on the measure's performance rate. Data used were for the time period described in Question 1.3.

- Risk adjustment: Not applicable; this measure is not risk adjusted.
- Meaningful difference in performance: We used electronically extracted EHR data from 11 hospitals to identify difference in performance by test sites and by demographic characteristics such as age, race, gender, sex, and payer source. Data used were for the time period described in Question 1.3.
- Missing data/bias: We used electronically extracted EHR data from 11 hospitals to examine the extent to which age and admission and discharge dates were missing in the electronically extracted data from test sites' EHR. Data used were for the time period described in Question 1.3.

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

As described in section 1.6, we collected information on the following variables using data extracted from hospital EHR systems: age, sex, race, ethnicity, and payer. This measure is based on a process that should be carried out for all patients (except those excluded), so no adjustment for patient mix is necessary. We did collect information about these five variables and assessed disparities in performance rate for each group. Those results are described in section 2b5.

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*)

We employed the split-half correlation to assess the reliability of the performance measure scores. The split-half correlation is a way to implement the test-retest reliability method. It estimates the measure reliability directly from the data and is less constrained by a small number of test sites than other model-based methods that require more data to justify model assumptions (for example, signal-to-noise using Beta-binomial model). The split-half correlation characterizes the correlation of estimated measure results between two non-overlapping data sets. To estimate the reliability, we randomly divided the hospital-level EHR data into two equal samples. We then calculated the measure performance in both samples for each hospital and calculated the weighted correlation between the estimates of the performance rate (the hospital's weight is based upon its number of denominator cases to account for the sample size effect in each hospital). The higher the correlation, the higher the statistical reliability of the measure. Stated another way, the higher the correlation, the greater the amount of variation that can be explained through systematic differences across the test sites as opposed to random error (for example, sampling variation within measured entities). To produce more stable estimates, we repeated this resampling approach more than 2,500 times. We used 0.4 as our benchmark level for an acceptable estimate of measure reliability because it aligns with guidance in the literature; Evans (1996) suggests that for the absolute value of Pearson's correlation *r*, a range of 0.40–0.59 indicates "moderate" reliability.

[Reference: Evans, J. D. (1996.) *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Publishing, Pacific Grove.]

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*)

Reliability tests were conducted, as described in section 2a2.2, to generate a reliability score for the measure. Because we are looking at measure-level reliability, the measure has one reliability score:

Table 3. Reliability testing results

Measure name	Reliability score	95% Confidence Interval
Use of Antipsychotics in Older Adults in the Inpatient Hospital Setting, 65 years of age and older	0.981	0.957, 0.995

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?*)

We assessed performance rate reliability across test sites using a split-half correlation. The reliability coefficient across 11 hospitals for the antipsychotic measure was .981 (with a 95 percent confidence interval, (0.957, 0.995) for all encounters, 65 years of age and older. This indicates that the hospital-level performance rate has excellent reliability, and is relatively free from measurement error. Reliability coefficients of .9 or above reflect excellent precision between performance rates derived from the two samples (a reliability coefficient of 1.0 reflects perfect precision).

[Reference: Adams, John L. "The Reliability of Provider Profiling: A tutorial." Santa Monica, CA: RAND Corporation, 2009.]

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ **Performance measure score**

☐ **Empirical validity testing**

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

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

Data element (Criterion) Validity

Data element validity testing evaluated whether the measure specification correctly identifies all the data elements required to calculate the measure score. This method quantifies the percent agreement, Kappa statistic, sensitivity, specificity, negative predictive value and positive predictive value between the electronically extracted EHR data and the manually abstracted data (which use the entire record, including free text notes fields). Each of these statistics illustrates the closeness between data element results from the two sources. In general, the higher the value, the more consistency between the data from the two sources.

Data element validity was tested by selecting a random set of patient encounters from the full electronic EHR extract and comparing data for these encounters to those that were manually abstracted, by trained abstractors, for the same encounters. The manually abstracted EHR data were considered the 'gold standard' against which we assessed the validity of the EHR-extracted data.

Face Validity

Formal Assessment of Face Validity (EWG and staff at test sites)

We evaluated the face validity of the measure specification and the measure score by surveying eight experts via the web: two clinicians from Test Sites 1 and 3, four Expert Work Group (EWG) members (three physicians, one academic), and two quality improvement / informatics staff from Test Sites 1 and 2. The survey asked respondents about the

appropriateness of the measure components (denominator, denominator exclusions, numerator, and numerator exclusions) given the intent of this measure. In addition, we asked respondents if hospitals that 1) document "threat of harm" for patients that are prescribed antipsychotics, and 2) document denominator exclusions, should score well on the measure. For each item, respondents indicated the extent to which they agreed (1 = Strongly agree; 2 = Agree, 3 = Disagree; 4 = Strongly disagree).

The EWG, which included physicians, academicians, and subject matter experts, helped ensure that the measure specification and measure score have a high degree of face validity. EWG members are listed in Table 4. We also evaluated the face validity of the measure specification and the measure score by soliciting input from key stakeholders during public comment, the patient and family advisory board (PFAB) and the Technical Expert Panel.

Table 4. EWG Members

Name	Organization
Byron Bair, MD, MBA	Salt Lake City VA
Soo Borson, MD	University of Washington
Josh Chodosh, MD, MSHS	NYU School of Medicine
Elizabeth Galik, RN, PhD, CRNP	University of Maryland School of Nursing
Susan Merel, MD	University of Washington Department of Medicine
Paul Rosenberg, MD	Johns Hopkins
Lynn Shell, PhD, APN, CARN-AP	Rutgers
Teepa Snow, MS, OTR/L, FAOTA	Positive Approach, LLC
Heidi Wald, MD, MS, MS	University of Colorado

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

Data Element Validity

There were high levels of agreement between the electronically extracted and manually abstracted EHR data for the denominator, denominator exclusions, numerator, and numerator exclusions across the three test sites. Table 5 describes the level of agreement between the two data sources for each component of the measure specification. The chart-abstracted data represent the gold standard for data element validity testing.

Table 5. Agreement statistics for random sample data between EHR extraction and manual chart abstraction (n=158)

Measure Component	Agreement (%)	Kappa	Sensitivity	Specificity
Denominator	98.1	0*	1	0
Initial Population	98.1	0*	1	0
Denominator exclusion				
Schizophrenia	99.4	0.66	0.88	0.99
Huntington's	100.0	NaN	NaN	1
Bipolar	98.8	0.49	0.5	0.99
Tourette's	100.0	NaN	NaN	1
Numerator (antipsychotic order during encounter)	100.0	1.0	NaN	1
Numerator exclusion	98.1	0.39	0.33	0.99

Source: Data from 10/1/2013 to 9/30/2015 for Test Sites 1 and 2, and 10/1/2014 to 9/30/2015 for Test Site 3.
Notes: NaN: Not calculable because the denominator in the equation is equal to zero.
*All 158 cases were contained within the denominator from the EHR. Chart abstractors flagged 3 of the 158 cases as not meeting denominator criteria. The Kappa statistic treats the 155 yes-yes agreement largely as "chance agreement" and penalizes this condition when applying the chance correction.

We measured overall agreement, defined as the number of patients for which both sources agree on the presence or absence of a condition among all patients tested. We also used Cohen's Kappa statistic to reflect chance-adjusted agreement. The Kappa score can range from -1.00 to 1.00. Although higher Kappa scores tend to indicate higher agreement between two data sources, a low Kappa score may not represent low agreement when the data are imbalanced.

The overall sample of 158 encounters showed 98 percent agreement or higher for all data elements and data element combinations assessed. In addition, agreement was perfect for two of the exclusionary data elements (Tourette's and Huntington's) and the numerator data element (antipsychotic prescription) and almost perfect for the remaining data elements. Kappa values ranged from a low of .39 for the numerator exclusion ("threat of harm") to a high of 1.0 for the numerator (medication orders). The numerator exclusion sensitivity is reflective of the inconsistent documentation of the numerator exclusion ("threat of harm") in the EHR.

[Reference: Viera, Anthony J., and Joanne M. Garrett. "Understanding Interobserver Agreement: The Kappa Statistic." Family Medicine, vol. 37, no.5, 2005, pp. 360–363.]

Face Validity

Results from the web-based survey of members of the measure's EWG, and test site representatives indicate that the measure had strong face validity. All respondents (n=8) strongly agreed or agreed that the measure components (denominator, denominator exclusions, numerator, and numerator exclusions) were appropriate to the intent of this measure (stated at the beginning of the executive summary). Further, six out of eight respondents agreed that hospitals should score well on the measure if they 1) document "threat of harm" for patients that are prescribed antipsychotics, and 2) document denominator exclusions.

Twenty-two comments were received during the 30-day public comment period which ran from April 15, 2016 through May 15, 2016. Commenters included hospitals and health systems (6), professional associations (7), EHR vendors (2), academic institutions (3), and individuals (2). Responses reinforced the measure's main goal of calling attention to off-label antipsychotic prescribing practices, thereby reducing inappropriate use of antipsychotics. Many commenters acknowledged the importance of developing a hospital measure that addresses use of antipsychotics in the inpatient setting. Some, however, highlighted the potential unintended consequences of the measure's implementation in critical care settings. Some of the commenters expressed concern over the overall intent of the measure, suggesting that the measure might unintentionally encourage the use of potentially harmful and less effective alternatives such as benzodiazepines. Also cited was the potential for increased use of physical restraints as an alternative to antipsychotics.

These commenters also questioned the ability of the measure to address either appropriate use of antipsychotics or quality care gaps.

The PFAB believed that the antipsychotic measure is important and that it could be used to help decrease the use of antipsychotics during hospitalization and, possibly, long-term. They believed the measure may facilitate proactive provider education and improved hospital policies on managing patient agitation. In addition, the measure may result in greater levels of engagement with the patient as well as his/her family.

The TEP was in agreement about the importance of the measure. There was concern about the measure being focused on medications ordered rather than medications administered. The intent of the measure is to change prescribing behaviors. In the future, CMS may consider adding a second numerator for antipsychotics administered.

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

The Kappa values calculated through data element validity testing suggest that data in the EHR accurately reflect patient care. In addition, face validity appears to be high as well. Six out of eight respondents reported that hospitals would score well on the measure if they consistently documented "threat of harm" and denominator exclusions. One person who disagreed commented that the denominator exclusions should be broader, noting that there were other diagnoses for which patients were on chronic antipsychotics. The other respondent who disagreed did not provide qualitative feedback.

_____

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b3

The following five exclusions apply to the measure. Excluded from the denominator are encounters with a documented diagnosis of schizophrenia, Huntington's, bipolar disorder, or Tourette's. These are conditions for which antipsychotics are approved for use. Excluded from the numerator are encounters with a documented "threat of harm to self or others." This exclusion is supported by clinical guidelines.

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

To examine the effect of the exclusions, the number affected by exclusions was first examined and the measure rates with and without each exclusion were calculated and compared.

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 6 shows the count of cases within each exclusion category across sites and within sites. It should be noted that within an encounter, a patient may have more than one exclusion.

Table 6. Number and proportion of exclusions

	Test site 1		Test site 2		Test site 3		Across sites (pooled data)	
	N	%	N	%	N	%	N	%
Number of encounters	45,097	100.0	12,954	100.0	456	100.0	58,507	100.0
Number of encounters in denominator exclusion	1,316	2.9	310	2.4	1	0.2	1,627	2.8
Number of encounters in numerator exclusion	104	0.2	48	0.4	0	0.0	152	0.3

Table 7 shows performance rates by test site for the measure as it is currently specified with exclusions (Column A), the measure with no numerator exclusion (Column B), the measure with no denominator exclusions included in the

calculation (Column C), and the measure including one of the four denominator exclusions (schizophrenia, Huntington's Disease, bipolar disorder, Tourette's Syndrome) (Columns D, E, F, G, respectively).

Table 7. Comparison of performance rate based on exclusion criteria

	COLUMN A	COLUMN B	COLUMN C	COLUMN D	COLUMN E	COLUMN F	COLUMN G
	Performance rate. Measure as specified	Performance rate without numerator exclusion	Performance rate without denominator exclusions	Performance rate including schizophrenia; excluding Huntington's, bipolar, and Tourette's	Performance rate including Huntington's; excluding schizophrenia, bipolar, and Tourette's	Performance rate including bipolar; excluding schizophrenia, Huntington's, and Tourette's	Performance rate including Tourette's; excluding schizophrenia, Huntington's, and bipolar
Total	21.6	21.7	22.6	22.0	22.5	21.9	22.6
Test site 1	21.5	22.7	22.5	22.0	22.5	21.8	22.5
Test site 2	22.5	22.7	23.4	22.9	23.4	23.0	23.4
Test site 3	6.6	6.6	6.8	6.8	6.8	6.6	6.8

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Performance rates vary little regardless of the denominator or numerator exclusions. When including all patients 65 years of age and older in the denominator regardless of diagnosis, the performance rate increases one percentage point from 21.6 percent (measure as specified) to 22.6 percent. Similarly, if we remove the numerator exclusion and include all patients who received an order for antipsychotics in the measure calculation regardless of "threat of harm" documentation, the rate increases slightly from 21.6 to 21.7 percent. This minimal difference is not surprising since it has been reported that "threat of harm" documentation is often lacking. Based on testing, the results suggest that numerator and denominator exclusions have little impact on the performance rate. However, for face validity, clinician acceptance of the measure, and consistency with clinical guidelines, it is recommended that the measure exclusions remain as specified.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with _risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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

Not applicable.

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

Not applicable.

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

Not applicable.

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

Not applicable.

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

Not applicable.

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

Not applicable.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

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

Not applicable.

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

Not applicable.

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

Not applicable.

2b3.9. Results of Risk Stratification Analysis:

Not applicable.

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

Not applicable.

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

Not applicable.

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

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

We analyzed the data to determine if there were statistically significant differences in performance rates by hospital or by age, sex, race, ethnicity, or payer. We also examined differences in performance rates based on intensive care unit (ICU) exposure (encounters with an ICU exposure vs. encounters without an ICU exposure).

To identify statistically significant differences in performance across multiple hospitals, we examined the distribution of performance rates across hospitals. In addition, we calculated the 95 percent confidence interval of the performance rate for each hospital using a z-distribution for proportion. Then we compared each hospital's confidence interval to the overall performance rate, which includes all patients across hospitals. Hospitals with confidence intervals higher than the overall rate indicate room for improvement.

In addition, we conducted chi-square tests to test statistically significant differences in performance between disparity groups, and between care settings.

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

Performance by hospital

In Table 8, we provide performance rates for each hospital across the three test sites. Performance rates varied from a low of 6.6 percent at Test Site 3 to a high of 25.9 at one of the hospitals in Test Site 1.

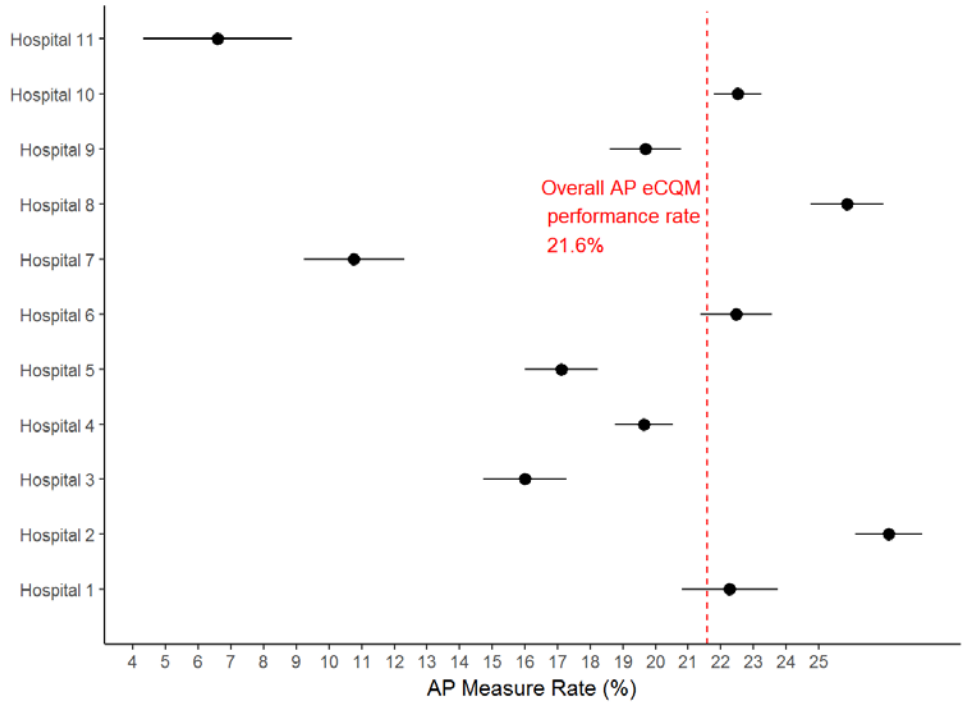
Table 8. Antipsychotic electronic clinical quality performance rates (hospital level)

Hospital	Antipsychotic order (%)
Test site 1	
Hospital 1	22.3
Hospital 2	27.1
Hospital 3	16.0
Hospital 4	19.6
Hospital 5	17.1
Hospital 6	22.5
Hospital 7	10.8
Hospital 8	25.9
Hospital 9	19.7
Test site 2	22.5
Test site 3	6.6

The variation in hospital-level measure performance is further illustrated in Figure 1, which shows the distribution of the performance rate and 95 percent confidence interval for each hospital relative to the overall performance rate. The confidence interval is the range in which each hospital's performance rate would likely fall if extractions were repeated multiple times. Although some hospital rates were below the overall performance rate, five out of the 11 hospitals (45.4

percent) have measure rates significantly higher than the overall measure rate (21.6%), indicating room for improvement.

Figure 1. Distribution of performance rates by hospital (orders)



Performance by disparity group

Age. The American Geriatrics Society 2015 Updated Beers Criteria for Potentially Inappropriate Medication Use in Older Adults cautions against the use of antipsychotics in patients age 65 and older.[1] Testing results indicated that patients ages 65 years and older were more likely to be ordered antipsychotics than patients ages 18 to 64 years , 21.6 percent versus 14.8 percent (not shown in table) (p<.001). Further, when limiting the analysis to those 65 years of age and older, we see a linear relationship with performance rates increasing as age increases (not shown). As seen in Table 9, among patients ages 65 to 74 years, 15.9 percent received an order for antipsychotics compared to 33.0 percent of patients ages 85 years and older (p<.001). This is an important finding as there has been significant concern about the inappropriate use of antipsychotics among older individuals. These findings support the notion that older patients are more likely to receive antipsychotics than younger patients, lending support to the importance of this measure.

Sex, race, and ethnicity. As seen in Table 9, males had higher rates of antipsychotic ordering than females, 24.0 and 19.7 percent, respectively (X^2=154.7, p<.001). With regard to race, Blacks were more likely than Whites to be ordered an antipsychotic, 24.4versus 20.9 percent, respectively (X^2=54.8, p<.001). There was little difference in the rate of antipsychotic ordering by ethnicity. Hispanic and non-Hispanic patients had similar performance rates, 20.6 and 22.0 percent, respectively (X^2=31.6, p<.001). Although differences based on sex, race, and ethnicity are small and likely not clinically significant, they are statistically significant. This is likely due to the large sample size.

Payer. Patients with Medicare and Medicaid coverage had the highest rates of antipsychotic ordering, 22.0 and 27.9 percent, respectively. Patients with private insurance had the lowest rate at 13.4 percent; this was expected as the measure is focused on older adults (65 years and older). Results were statistically significant (X^2=161.2, p<.001).

[1] The American Geriatrics Society (AGS) Beers Criteria for Potentially Inappropriate Medication (PIM) Use in Older Adults is an explicit list of PIMs best avoided in older adults in general and in those with certain diseases or syndromes, prescribed at reduced dosage or with caution or carefully monitored. It is one of the most frequently consulted sources about the safety of prescribing medications for older adults. The AGS Beers Criteria are used widely in geriatric clinical care, education, and research and in development of quality indicators. Accessed on June 21, 2017 at https://guideline.gov/summaries/summary/49933/american-geriatrics-society-2015-updated-beerscriteria-for-potentially-inappropriate-medication-use-in-older-adults?q=diabetes.

Table 9. Performance rate for antipsychotic ordered measure by patient characteristic

Characteristics	Test site 1		Test site 2		Test site 3		Across sites (pooled data)	
	N	Performance Rate (%)	N	Performance Rate (%)	N	Performance Rate (%)	N	Performance Rate (%)
Number of patients	45,097		12,954		456		58,507	21.6
Average age	77.0		74.6		78.1		76.5	
Age								
65 to 74	20,178	14.7	7,414	19.4	191	7.3	27,783	15.9
75 to 84	15,201	22.6	3,963	24.2	143	3.5	19,307	22.8
85 and older	9,718	33.4	1,577	32.4	122	9.1	11,417	33.0
Sex								
Male	18,948	23.8	6,277	25.1	172	5.2	25397	24.0
Female	26,145	19.8	6,677	20.1	284	7.4	33,106	19.7
Race								
White	28,381	20.8	9,810	21.7	455	6.6	38,646	20.9
Black	6,407	24.1	2,543	25.1	1	0.0	8,951	24.4
Other	8,439	22.5	457	25.1	0		8,896	22.7
Ethnicity								
Hispanic	4,889	20.7	175	18.8	0		5064	20.6
Non-Hispanic	37,215	21.9	12.404	22.5	436	6.9	50,055	22.0
Other or unknown	2,993	16.8	375	24.1	20	0.0	3,388	17.7
(Primary) Payer								
Medicare	41,415	22.0	11,924	22.5	279	6.8	53,618	22.0
Medicaid	585	28.7	83	25.3	9		677	27.9
Private insurance	2,513	11.9	764	20.2	166	6.7	3,443	13.4
Self-pay or uninsured	283	19.4	63	21.7	0		346	19.8
Others	301	24.0	120	32.8	2	0.0	423	26.4

SOURCE: Test Site 1 and Test Site 2 data from October 1, 2013 to September 30, 2015. Test Site 3 from October 1, 2014 to September 30, 2015.

Performance by ICU exposure

We examined the rate at which antipsychotics are ordered in the ICU as compared to non-ICU settings. With the data available, we were able to look at this in two ways. Using method 1, we determined the unit where the patient was *first* assigned at the time of admission. If the patient was assigned to the ICU, the patient's encounter was classified as ICU. Other encounters were classified as non-ICU. Using method 2, we assigned patients to the ICU group if they were assigned to the ICU at *any* point in time during their encounter.

Both methods, as seen in Table 10, yield similar rates of antipsychotic ordering. Using method 1, 37.5 percent of ICU patients were ordered an antipsychotic during their encounter as compared to 21.4 percent of non-ICU patients.

Using method 2, 37.7 percent of ICU encounters and 27.9 percent of non-ICU encounters had an antipsychotic order. Differences by ICU exposure were statistically significant, across both methods (X^2=87.7, p<.001).

Table 10. ICU versus non-ICU performance rates for antipsychotic ordering

	Method 1		Method 2	
	ICU	Non-ICU	ICU	Non-ICU
Performance rate	37.5	21.4	37.7	27.9

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

The results demonstrate that statistically significant differences can be detected between hospitals. The variations in performance across hospitals suggested meaningful differences in the quality of care provided between the lowest and highest performing hospitals and indicated that there is ample room for improvement. In addition, disparities in performance based on age, sex, race, ethnicity, and payer further suggested room for improvement. The statistically significant difference in antipsychotic ordering in the ICU versus non-ICU settings encouraged us to include stratification by unit of care in the measure specification.

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

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

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

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

Not applicable.

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

Not applicable.

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

Not applicable.

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences

between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Date of birth is required for the measure calculation, as it is applicable for patients ages 65 years and older. In addition, encounters are defined by admission and discharge dates. Missing data on date of birth and admission and discharge dates was negligible. Missing data is not a threat to validity for this measure. The majority of data elements required to calculate the performance rate are ones in which absence of data in a data field reflects the absence of a condition or behavior (for example, diagnosis or medication ordered).

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

See response for 2b6.1

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

See response for 2b6.1

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)

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*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in electronic health records (EHRs)

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

Not applicable

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. Please also complete and attach the NQF Feasibility Score Card.

Attachment: AP_Feasibility.xlsx

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

Not applicable

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)*.

Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.

Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html).

There are no other fees or licensing requirements to use this measure, which is in the public domain.

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.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Payment Program Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

Not applicable; the measure is under initial endorsement review and is not currently used in an accountability program.

4a1.2. 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?*)

CMS is considering implementation plans for this measure. There are no identified barriers to implementation in a public reporting or accountability application.

4a1.3. 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 measure has been submitted through the Measures Under Consideration process for the CMS Hospital Inpatient Quality Reporting Program and Medicare and Medicaid Programs; Electronic Health Record Incentive Program – Stage 3.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Not applicable

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Not applicable

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Not applicable

4a2.2.2. Summarize the feedback obtained from those being measured.

Not applicable

4a2.2.3. Summarize the feedback obtained from other users

Not applicable

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Not applicable

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

Adoption of this performance measure has the potential to improve the quality of care for hospitalized older adults in the area of patient safety, a priority area identified by the National Quality Strategy. Specifically, this measure will encourage thoughtful prescribing of antipsychotics for hospitalized patients and an increase in non-pharmacologic treatments. More careful antipsychotic prescribing among hospitalized individuals would be expected to result in fewer prescriptions continued after discharge and, ultimately, lower morbidity and mortality associated with the long-term use of these medications.

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

This measure has not been implemented. During measure testing, experts suggested that a potential unintended consequence could be the increased use of alternative harmful medications such as benzodiazepines for delirium or BPSD. Additionally, clinical guidelines recognize that pharmacologic options should be a last resort after careful consideration and only after nonpharmacologic interventions have failed. This suggests that off-label antipsychotic use in the inpatient setting should not be expected to reach zero as clinical judgment will need to be exercised in situations where an alternative treatment may not be available or address the specific patient circumstances.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

This measure has not been implemented. During measure testing, experts suggested that a potential benefit could be more thoughtful prescribing of antipsychotics in the inpatient setting, as well as fewer continued prescriptions after discharge to other care settings. This could encourage the use of delirium assessment and monitoring tools, improved detection of patient behaviors that could otherwise escalate to delirium, and the use of nonpharmacologic interventions to manage behavior.

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.

Yes

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

2111 : Antipsychotic Use in Persons with Dementia

2993 : Potentially Harmful Drug-Disease Interactions in the Elderly

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

The following are related measures not currently endorsed by NQF:

- CMS N011.01: Percentage of [Nursing Home] Residents Who Newly Received an Antipsychotic Medication (Short Stay)
- CMS N031.02: Percentage of [Nursing Home] Residents Who Received an Antipsychotic Medication (Long Stay)

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

Yes

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

These measures are harmonized to the extent possible. While all measures assess the potentially inappropriate use of antipsychotic medications, this is the only measure that assesses use of antipsychotic medications in the inpatient hospital setting. CMS N011.01 and CMS N031.02 are intended for use in the nursing home setting. Measures NQF 2111 and NQF 2993 assess health plan performance. This measure's eligible population includes all patients in an inpatient hospital setting who are age 65 and older, which aligns with the age for measures NQF 2111 and NQF 2993. NQF 2111 and NFQ 2993 only assess older adults with dementia, whereas this measure includes all older adults. The denominator exclusions are similar across measures. The exclusions in this measure—schizophrenia, Tourette's syndrome, Huntington's disease, and bipolar disorder—are similar to exclusions in related measures. CMS N011.01, CMS N031.02, and NQF 2111 exclude patients with schizophrenia, Tourette's syndrome, or Huntington's disease. NQF 2111 also excludes patients with bipolar disorder. NQF 2993 excludes patients with psychosis, schizophrenia, or bipolar disorder. This measure also excludes from the numerator people in the inpatient setting who are identified as a threat to themselves or others. No other measure excludes these patients, although this exclusion is appropriate for the hospital setting. The specific antipsychotic medications included in each measure are the same, with only three exceptions; NQF 2111 does not include brexpiprazole, cariprazine, and molindone whereas NQF 2993 and the measure under development include these medications.

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.)

Not applicable

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: AP_LogicFlow_for_S.14_response.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Joseph, Clift, joseph.clift@cms.hhs.gov, 410-786-4165-

Co.3 Measure Developer if different from Measure Steward: Mathematica Policy Research

Co.4 Point of Contact: Brenna, Rabel, brabel@mathematica-mpr.com, 609-945-6564-

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.

Antipsychotics Measure Development Expert Work Group:

This panel provided expertise in geriatric and inpatient care and provided feedback on the measure specifications and testing results.

--Byron Bair, MD, MBA - Salt Lake City VA

--Soo Borson, MD - University of Washington

--Josh Chodosh, MD MSHS - NYU School of Medicine

--Elizabeth Galik, RN, PhD, CRNP - University of Maryland School of Nursing

--Susan Merel, MD - University of Washington Department of Medicine

--Paul Rosenberg, MD - Johns Hopkins University

--Lynn Shell, PhD, APN, CARN-AP – Rutgers University

--Teepa Snow, MS, OTR/L, FAOTA - Positive Approach, LLC

--Heidi Wald, MD, MS, MS - University of Colorado

Technical Expert Panel:

This panel provided overall guidance on measure development and project direction, including review of the measure specification and testing results.

--Peter Bach, MD, MAPP, Memorial Sloan Kettering

--James Burgess, PhD (co-chair) - Boston University

--Donna Slosburg, RN, BSN - ASC Quality Collaborative

--Ileana Pina, MD, MPH - Albert Einstein College of Medicine

--Jeremiah Schuur, MD, MHS - Brigham and Women's Hospital

--John Hertig, PharmD, MS - Purdue University

--Marc Overhage, PhD, MD - Siemens Health Services

--Kent Sepkowitz, MD - Memorial Sloan Kettering Cancer Center

--Maureen Dailey, PhD, RN - American Nurses Association

--Michael Howell, MD, MPH (chair) - University of Chicago Medicine

--Monica Peek, MD, MPH - Chicago Center for Diabetes Translation Research

--Nancy Foster - American Hospital Association

--Nathan Goldstein, MD - Mount Sinai School of Medicine

--Stephen Edge, MD - Baptist Cancer Center

--Susan McBride, PhD, RN-BC - Texas Tech University Health Sciences Center

--Thomas Louis, PhD - Johns Hopkins Bloomberg School of Public Health

Patient and Family Advisory Board:

This panel provided feedback on the measure concept from the patient and family perspective.

--Darlene Barkman - Children's Hospital of Philadelphia

--Ann Cannarozzo - Rochester Regional Health System

--Maureen Corcoran - Cystic Fibrosis Foundation

--Ilene Corina - PULSE (Persons United Limiting Substandards and Errors in Healthcare) of NY

--John Harris - Johns Hopkins Hospital

--Toby Levin - Suburban Hospital Patient and Family Advisory Council

--Christopher Mason - Peace Health Patient Advisory Council

--Teresa Masters - Patient and Family Centered Council, University of California, San Diego

--Lisa McDermott - National Brain Tumor Society

--Kelly Parent - Patient and Family Centered Care Program, University of Michigan Health System

--Lee Tomlinson - Center for More Compassionate Care

--Karel Shapiro - Rochester General Hospital

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? Specifications for this eCQM will be reviewed and updated annually.

Ad.5 When is the next scheduled review/update for this measure? 12, 2018

Ad.6 Copyright statement: Limited proprietary coding is contained in the Measure specifications for user convenience. Users of proprietary code sets should obtain all necessary licenses from the owners of the code sets.

Ad.7 Disclaimers: These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 Additional Information/Comments: Not applicable

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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3332

Measure Title: Psychosocial Screening Using the Pediatric Symptom Checklist-Tool (PSC-Tool)

Measure Steward: Massachusetts General Hospital

Brief Description of Measure: Percentage of children from 3.00 to 17.99 years of age seen for a pediatric well child visit who have a Pediatric Symptom Checklist (PSC) Tool administered as a component of that visit.

Developer Rationale: Psychosocial problems in children are common and treatable with prevalence estimates of about 12% of all children and adolescents (Gardner, Lucas, Kolko, & Campo, 2007; Kelleher et al., 1997; Murphy et al., 2016). Studies have shown that children with these problems are often unrecognized by their pediatricians (~50% of cases) (Kelleher et al., 1997) and that only a fraction of them receive treatment (Hacker et al., 2014b; Kelleher et al., 1997). Children with psychosocial problems are more likely to have poorer health, academic, behavioral, and social outcomes in both the short and long term (Murphy et al., 2015). Children who receive psychosocial screening as a part of pediatric well child visits are more likely to receive outpatient mental health services (Hacker et al., 2014a; Hacker et al., 2014b; Savageau et al., 2016) than are children who are not screened. As the dates of the studies just cited attest, it is only within the last three years that strong evidence documenting the relationship between psychosocial screening and increased mental health treatment has become available.

A series of RCT studies by Kolko and his associates have shown that pediatric outpatients with a wide range of problems who are found to be at risk when screened with the PSC and go on to receive pediatric office based mental health interventions have significantly lower mental health symptom scores and better functioning at immediate and longer term follow up than do similar outpatients randomized to treatment as usual (Kolko et al., 2014; Kolko, Campo, Kelleher, & Cheng, 2010). For these reasons, we believe that an increase in mental health treatment is the most appropriate (and a measurable) benchmark for assessing the positive impact of routine psychosocial screening. The logic model for screening in pediatrics is that more children will receive help, fewer children will develop mental, emotional, and behavioral disorders (Guzmán et al., 2015; Kieling et al., 2011), and more children who received help will enjoy better life outcomes (Kellam et al., 2014).

Requiring screening for psychosocial problems as part of routine well child care in pediatrics is one of the most frequently recommended ways to improve recognition and intervention for such problems (Hacker et al., 2014a) and an increasing number of states (Massachusetts (Savageau et al., 2016)), insurers (Medicaid/EPSDT (Mann, 2013)), standard setting organizations (American Academy of Pediatrics (Foy, Kelleher, Laraque, & Health, 2010; Weitzman & Wegner, 2015)), blue ribbon panels (President's New Freedom Commission on Mental Health (Hogan, 2003) (Institute of Medicine (O´Connell, Boat, & Warner, 2009)), and advocacy organizations such as the Kennedy Forum (Fortney et al., 2015) and Mental Health America (http://www.mentalhealthamerica.net/positions/early-identification) have now

required, endorsed, or recommended the principle of including a psychosocial screen as a part of every well child visit for children aged 3-17.

The PSC is probably the most frequently recommended and widely used tool for routine psychosocial screening in pediatrics (Semansky, Koyanagi, & Vandivort-Warren, 2003), with the Strengths and Difficulties Questionnaire (Goodman, Meltzer, & Bailey, 1998) and Child Behavior Checklist (Achenbach, 2009) instruments that are similar in many ways and also frequently mentioned and validated in this context. Many of the endorsements noted above include these three and/or a few other instruments.

The reference list is included in the attached appendix.

Numerator Statement: Number of patients with documentation that the PSC tool was administered as part of the well child visit.

Denominator Statement: Number of patients aged 3.00 to 17.99 seen for a pediatric well-child visit.

Denominator Exclusions: No exclusions.

Measure Type: Process

Data Source: Claims, Electronic Health Records, Paper Medical Records

Level of Analysis: Clinician : Group/Practice, Facility, Population : Regional and State

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

Staff Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* 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. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

The developer provides a logic model indicating that the process measure will increase the likelihood that psychosocial issues have been identified, explored and dealt with during a well-child visit.
The developer presents the following Evidence to support this measure:

- More than 180 studies over the past 30 years demonstrating the feasibility and acceptability of the PSC as a clinical and research measure with diverse populations.
- Studies demonstrating feasibility on a statewide scale and sustainable over nearly a decade.
- Strong evidence that children who have a positive risk score on the PSC are more likely to be referred to and to receive mental health services.

- Professional and advocacy organizations (e.g. American Academy of Pediatrics, the president's New Freedom Commission on Mental Health, and Institute of Medicine) have published recommendations in support of the measure.
- Guidelines governing the U.S. Medicaid program, Early and Periodic Screening, Diagnostic and Treatment (EPSDT), require routine mental health screening as part of a well-child visit (WCV).

Exception to evidence
N/A

Questions for the Committee:

- *How strong is the evidence for the relationship of this measure to patient outcomes?*
- *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm: Process measure with empirical evidence submitted but not systematically reviewed (Box 3) → empirical evidence without systematic review/grading of evidence (Box 7) → evidence summarized include all studies (box 8) → high certainty (box 9) → Moderate
The highest rating possible is MODERATE.

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities Data

Maintenance measures – increased emphasis on gap and variation

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

The developer provides data from several studies to demonstrate the wide variation in the rates of mental health screening with formal, standardized tools. The data presented is based on data from the Massachusetts Medicaid (MassHealth) pediatric behavioral health screening program and the Children's Behavioral health Initiative (CBHI).

- Measurement Period dates of service from 1/1/2008 to 3/31/2017. Data Source: Statewide data for all children

Denominator/Well child visits for all children .5 -20 years of age:	**4,721,790**
Numerator (screens with visit):	2,965,923
Statewide average:	62.8%
Minimum:	14.2%
Maximum:	71.9%
Standard Deviation:	12.4%
95% Confidence interval:	58.8% to 66.8%

- Below presents the distribution of rates of screening for all 222 measurement points (37 quarters in all six regions). Measurement Period dates of service from 1/1/2008 to 3/31/2017. Data Source: Statewide data broken down by for children ages 3-17 years

Denominator/Well child visits for all children 3 -17 years of age:	**2,361,475**
Numerator (screens with visit):	1,681,764
Statewide average:	71.2%
Minimum:	39.98%

Denominator/Well child visits for all children 3 -17 years of age:	**2,361,475**
Maximum:	79.14%
Standard Deviation:	11.3%
95% Confidence interval:	64.2% to 78.2%

- Massachusetts statewide number of well child visits, number of screens, and percent of visits with screens, for just the 3-17 year old (PSC screened) children with Medicaid from January 2008 (start of CBHI) to March of 2017 rise from approximately 39.98% for its first year (2008) to 65.72% for its second year to over 70% for its third year, and then remaining in the 70% range in all of the six years since.

  Measurement Period dates of service in 2007, 2010, and 2012. Data Source 2: CBHI Cohort Data for chart review sample. This data source is a ~ 6000 visits subsample of the CBHI statewide data retrieved from chart reviews supplemented by administrative claims data.

Denominator/Well child visits for all children:	**4,977**
Numerator (screen at WCV):	~1,700
Average:	51.7%
Minimum:	1.5%
Maximum:	88.9%
Standard Deviation:	35.1%
95% Confidence interval:	35.5% to 87.2%

- Measurement Period dates of service from 7/1/2014 to 12/31/2016. Data Source 3: Medicaid screening data from four Massachusetts General Hospital outpatient clinics

Denominator/Well child visits for all children 4-17 years of age:	**10,334**
Numerator (screen at WCV):	7,915
Average:	76.6%
Minimum:	9.4%
Maximum:	91.7%
Standard Deviation:	38.5%
95% Confidence interval:	38.8% to 100.0%

- To demonstrate differences between different pediatric practices, the developer extracted data from four MGH-affiliated outpatient pediatric clinics.

  2016 (n=6,801 children): January 1 2016 – December 31 2016

Variable	**Overall***	**Clinic A**	**Clinic B**	**Clinic C**	**Clinic D**
Billed for MH screen	69.7% (4743)	86.1% (3422)	88.1% (1215)	0.0% (0)	24.1% (106)

Variable	Overall*	Clinic A	Clinic B	Clinic C	Clinic D
p<.001					

Disparities

- The developer presents data from the CBHI BHSCQR showing differences in rates of screening by age group with very young (<3) children and older (>17) youth less likely to be screened than 3-17 year olds.

  (December 31 2007 - March 31 2017)

Age	Total Visits	Total Visits with Screens	% Visits with Screens
<6mos to 2 yrs	2,136,135	1,205,607	56.44%
<6mos	882,43	336,179	38.10%
6mos to 2 yrs	1,253,701	869,428	69.35%
3 yrs to 17 yrs	2,361,475	1,681,764	71.22%
3 yrs to 6 yrs	776,570	562,311	72.41%
7 yrs to 12 yrs	911,693	666,266	73.08%
13 yrs to 17 yrs	673,212	453,187	67.32%
18 yrs to 20 yrs	224,180	78,552	35.04%
Total	4,721,790	2,965,923	62.80%

- The developer provides data from a chart review study exploring screening by demographics in a subsample of ~6000 visits from 2007, 2010 and 2012.The developer reports no significant disparities by race, ethnicity, or language.

	Had a Formal Screen in 2007	Had a Formal Screen in 2010	Had a Formal Screen in 2012
Race			
White	7 (2.9%)	374 (73.6%)	397 (73.7%)
Non-White	20 (6.4%)	276 (77.7%)	290 (79.2%)
Ethnicity			
Non-Hispanic	17 (4.4%)	321 (75.3%)	369 (78.7%)
Hispanic	10 (5.8%)	224 (79.4%)	246 (80.9%)
Primary Language			
English	24 (3.8%)	676 (75.4%)	716 (76.8%)
Non-English	10 (6.5%)	128 (71.1%)	149 (76.4%)

Questions for the Committee:

- *Specific questions on information provided for gap in care.*
- *Is there a gap in care that warrants a national performance measure?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence

Comments:

**First, I am not clear how to classify this measure. The PSC is a composite; it reflects reported (by proxy) patient symptoms/behaviors/signs. Yet, it is being treated as a process (screen not screen). I am not convinced this is appropriate or consistent. Second, as in so many screening processes, the real juice is obtained not with the performance of the screen, but in the follow up and influence of patient oriented outcomes. Just to say we did more counseling or uncovered more symptoms doesn't really matter to children and their parents. And in fact, such use of services and labeling may indeed carry significant harms (glossed over by the accompanying analysis). So the actual causal pathway here is pretty tenuous and I am not convinced many of the "issues" uncovered by screening are highly important.

**evidence is strong for use as a process measure. Since initial review of this measure there has been considerable growth in the evidence to support this measure.

**moderate evidence to suggest screening leads to treatment which leads to better outcomes.

**A series of RCT studies have shown that use of the PSC to assess risk and who receive pediatric office-based mental health interventions have significantly lower mental health symptoms and better functioning in the near and longer term than others. Improving recognition and intervention for psychosocial problems in children has very long term results. The PSC is the most commonly used screening tool. Large number of studies over the past 30 years using the PSC...statewide feasibility has been shown. Tool is in use.

**process measure; empirical evidence provided.

**Did not submit evidence supporting the PSC over other evidence-based screening tools

**Process measure listing evidence of literature indicating use of PSC tool in children is useful.

**The evidence that required use of the PSC leads to greater screening is strong. There is good evidence that required use of the PSC in Massachusetts resulted in receipt of greater BH treatment. However, there is weak evidence that universal screening improves outcomes. A recent study found no effect of mandated screening in Massachusetts on BH-related hospitalizations, ED visits, or psychotropic drug use. The only study cited in to support the causal connect are by Kolko et al,. 2011 and Kolko et al., 2014 and these are studies of collaborative care not the effect of the use of the PSC in typical pediatric practices.

**The evidence supports that wide spread screening improves recognition and some sort of pediatric mental health intervention rates. As I recall last time we reviewed this measure the developer was claiming that it improved the lives/functioning of children involved. Now it seems more appropriately limited to an outcome of getting high risk kids into a mental health intervention

**Agree with moderate rating for evidence

**There is considerable evidence that screening in childhood well visits offers opportunities for early identification and intervention of mental health problems. The PSC has wide evidence of its efficacy. While there are other standardized tools available, developers offer considerable evidence to support that PSC be the sole tool for this quality measure.

**PSC has considerable evidence and is a well constructed tool. (I don't think it should replace PHQ-9 measure though as PSC is not a robust enough suicide measure.)

**The evidence applies directly and comes from more than 180 studies demonstrating the acceptability and feasibility of the measure. There were statewide studies and professional and advocacy groups published recommendations in support of the measure. EPSDT requires mental health screening as part of a well child visit.

** They show evidence of the need and data from Mass.

1b. Performance Gap

Comments:

**Yes, there is differing performance. I am not clear how well the findings from the Mass experience transfer to other locales.

**Yes, performance gap demonstrated.

**there are variations in performance by region and (based on limited data) by practice. the evidence is dominated by information from Massachusetts which probably has the net effect of overestimating performance relative to other places.
**Performance gap in Massachusetts shows that over a 10 year period screening increased overall from 40% to about 70% statewide and in MGH to about 75% but remains stuck at that level. And among billing for pediatric clinics screening is highly variable. While this may be a "billing issue" it also suggests significant implementation problems. Disparities in screening are important---data show that 13-17 year olds are less likely to be screened which, to this reviewer, is an important performance issue.
**significant gap demonstrated; with improvements demonstrated between measurement times from 2008 - 2017.
**Strong evidence that pediatric behavioral health screening is not widespread.
**Developer provided data demonstrating varying rates of use of this tool.
**The developer demonstrates significant variation across regions in MA, MA general hospitals, outpatient pediatric clinics, and by race/ethnicity/primary language in the use of the PSC.
**Good evidence that there is room for improvement
**Agree with assessment of gap, also would emphasize the importance of getting treatment to minors
**Gap appears to be in use of the tool by providers and interpretation of results. However, there does not appear to be a gap for this particular tool such that performance would be hindered.
**The data demonstrates quality prob,EMSI and opportunities for improvement. The Medicaid data from 1/08-3/17 shows dramatic improvement in the use of the measure. There are no disparities although older children are less likely to be screened.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability; Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable): Scientific Acceptability Form

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*

Questions for the Committee regarding validity:

- *The staff is satisfied with the validity analyses for the measure. Does the Committee agree?*

Preliminary rating for reliability:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability

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.**

Measure Number: 3332

Measure Title: Psychosocial Screening Using the Pediatric Symptom Checklist-Tool (PSC-Tool)

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?
   NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

   TIPS: Consider the following: 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?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.

2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

   TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

   ☒Yes (go to Question #4)

   ☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

   The developer provided data element validity testing. Per NQF guidance, if a developer performs data element validity testing, then data element reliability testing is not required. The developer also evaluated the inter-rater reliability of the assessments.

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☒Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

☐Yes (go to Question #5)

☒No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☐Yes (go to Question #6)

   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation?

   Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☐High (go to Question #8)

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

   ☐Yes (go to Question #8)

   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☒Yes (go to Question #9)

   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

   ☒Yes (go to Question #10)

   ☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

   Analyses showed that that the presence of the PSC or other brief BH screen during WCV could be reliably coded from the presence of the CPT 96110 code in administrative claims data. This approach to coding was also validated by the finding that 100% of the WCV that were billed for were documented by EMR notes from the same day and that the age codes for the WCV were congruent with the age of each child.

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

    TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☐Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW)

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

   TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

   ☒Yes (go to Question #2)

   ☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

   TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

   ☐Yes (please explain below then go to Question #3)

   ☐No (go to Question #3)

   ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

   ☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

   a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

   b. Are social risk factors included in risk model? ☐Yes ☐No

   c. Any concerns regarding the risk-adjustment approach?

   *TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do*

you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?

☐Yes (please explain below then go to Question #4)

☐No (go to Question #4)

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

   ☐Yes (please explain below then go to Question #5)

   ☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

   ☐Yes (please explain below then go to Question #6)

   ☐No (go to Question #6)

   ☒Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?

   ☐Yes (please explain below then go to Question #7)

   ☒No (go to Question #7)

ASSESSMENT OF MEASURE TESTING

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

   ☒Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

   ☐No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

   ☐Yes (go to Question #9)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

   ☐Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☐Yes (go to Question #11)

    ☒No (please explain below and go to Question #13)

The developer provided data element validity.

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☐Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #14)

    ☐Moderate (go to Question #14)

    ☐Low (please explain below then go to Question #13)

    ☐Insufficient

13. Was other validity testing reported?

    ☒Yes (go to Question #14)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

    TIPS: Prior validity studies of the same data elements may be submitted

    ☒Yes (go to Question #15)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☒Yes (go to Question #16)

    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☒Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

Comments:

**The idea of dichotomizing (screened/not screened) obviates a much more thorny assessment of data level reliability. I am not sure I understand the rationale for this approach. To take a ludicrous example, imagine I come up with a survey of items such as your favorite color, shoe size, interest in Star Trek, and whether you worship Steve Bannon. Of course this measure of mental health might be a bit suspect. But then all I do is find out if my new screener was done or not. GIGO. This certainly applies to validity and I believe reliability at the data element level.

**clearly specified

**No concerns with specifications.

**The time frame for the measure is not clear. The numerator "Number of patients with documentation that the PSC tool was administered as part of the well child visit." The term "documentation" is not clear. does this refer to a note in the patient chart or a claim submitted? It appears this is a claim based measure.

**Data elements are clearly defined.

**It's reasonably reliable

**Good data for reliability of PSC as measure and administrative code reliably captured the screening

2a2. Reliability – Testing

Comments:

**No concerns.

**No concerns with testing.

**3-17 is a large age spread. There may be other screening tools more appropriate for adolescents like the CRAFFT. Also, a tool for adolescents that is self- administered might be more appropriate.

**PSC is valid as is use of the administrative code, however, since the administrative code isn't specific for the PSC, it's not clear to me why the measurement definition specifies PSC in the measure numerator if it's predicated on the non-specific administrative code.

2b1. Validity –Testing

2b4-7. Threats to Validity

2b4. Meaningful Differences

Comments:

**I am not sold. See above

**problems of missing data I assume would be addressed in the data analysis and mentioned during reporting

**They show important differences e.g., by region and by practice. No substantial concerns.
**No concerns.
**Often a claim is dropped but there is not note int he clinical file to match.
**missing data was marked as n/a, however, it seems likely the data sets may have had missing data?

2b2-3. Other Threats to Validity

2b2. Exclusions

2b3. Risk Adjustment:

Comments:

**Probably not, although one could imagine language disparities driving important differences is performance.
.** no risk adjustment required for this process measure
**There does not seem to be a need to risk adjust. No concerns.
**No concerns. It's interesting that there is not much of a racial gap
**The literature suggests that providers need training in how to discuss potential mental health problems/survey outcomes/next steps with parents and that could become apparent as more pediatricians use this tool - could expose gaps for working with certain populations or geographic areas. This should definitely be coupled with outcome measures regarding continuity of care.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

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 noted the following:

- All data elements are in defined fields in a combination of electronic sources. Data elements are generated or collected by and used by healthcare personnel during the provision of care, coded by someone other than the person obtaining original information, abstracted from a record by someone other than person obtaining original information.
- There are no fees, licenses or other requirements needed to use any aspect of the measure or the instrument.

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

Comments:

**It has been shown to be feasible when mandated, but not clear it is as easy when not mandated. But not a major concern.
**highly feasible
**elements are routinely collected and used.
**no issues with feasibility; electronic sources are used, coded by someone other than the person originally obtaining the info; abstracted from record. No fees.
**This depends upon the definition of "documentation"
**Some concern about manual collection of data. This is dependent upon the EMR used and how the tool information is put into the EMR.
**No concerns.
**The implementation shows an admirable increase in utilization in the real world over time.
**Agree with moderate feasibility rating
**This is feasible, though providers struggle with how to finance screening and you may have to provide a list of codes to ensure that providers understand reimbursements. This could easily be completed by parents in waiting room. Unclear if there are versions in other languages. This should not be exclusionary criteria and health care providers

should ensure mechanism to screen all patients regardless of language. Looks like the denominator includes only screens on same day as well child visit, may want to consider revising to include screens within 48 hours (or week) prior to allow for some systems to push this out through a patient portal.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact/improvement and unintended consequences

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- This measure is publically reported in the Behavioral Health (BH) Screening Cumulative Quarterly Report (geographical area is the state of Massachusetts. Accountable entities include all providers of WCV to MassHealth members).
- This measure is used for Professional Certification or Recognition Program: Program MOC Part 4 Certification.
- This measure is used for Quality improvement with Benchmarking (external benchmarking to multiple organizations).

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others N/A

Additional Feedback: N/A

Questions for the Committee:

Preliminary rating for Use: ☐ **Pass** ☒ **No Pass**

RATIONALE: Feedback unknown by developer. Therefore, not provided. This criterion is not a requirement for a new measure.

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

Improvement results:

- The developer provides data demonstrating the increase in rates of screening in the full sample of 3-17 year olds over the first 9.25 years of CBHI.
- The developer provides data demonstrating increases present in all regions of the state.
    - Rates of screening for all ages increased dramatically, 14-fold, from ~ 5% of all WCV to > 70% over the first three years of the program and have been sustained at that level ever since.
    - For 3-17 year olds who were screened primarily with the PSC, the number of visits screened rose steadily over the first 7 years of the program (from ~80,000 to over 218,000 per year and from 40% to 79%).
- Data from two state (California and Massachusetts) comparisons demonstrate BH screening had risen to 13 per thousand children in Massachusetts while remaining at the same level in California.

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation

The developer lists the following benefits of implementing this measure:

- Increasing widespread use of a simple but effective PRO tool that can be used for screening, diagnosis and the monitoring of treatment outcomes for psychosocial problems (Massachusetts and other states).
- Increased national use of the measure (PSC is being used in the SAMHSA National System of Care Expansion Evaluation), Mental Health America making the PSC and PSC-Y available for free and to tens of thousands of youth or their parents.

Potential harms: The developer reposts no unintended negative consequences to individuals or populations.

Additional Feedback:

N/A

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

Preliminary rating for Usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

**The developers have persevered and been responsive to earlier feedback from prior review.
**user feedback has not been collected but the measure has been used widely in one state.
**Publicly reported in Massachusetts and used for professional certification and is used for quality improvement and benchmarking by many organizations.
**no feedback provided.
**The screening would be more useful if we also knew i the number of pediatric patients that received treatment.
**Has been widely used.
**The feedback loops and their impact are not clear to me. The fact that rates have improved probably indirectly reflect that this is occurring working
**The use and results of the measure have been publically reported in the Behavioral Health Screening Cumulative Quarterly Report. It is used for Professional Certification in Recognition Program and for QI benchmarking.

4b1. Usability – Improvement

Comments:

**The authors skip any hint of harms, yet we know screening can often bring up issues of stigma, inappropriate medication, increased costs of care, etc. They just don't seem to have looked for it.
**Early detection of probable psychosocial problems in pediatric settings during well child visits outweighs any risk for false positive of child psychopathology given the content, how the PSC is scored, and well established wording of what the cut-points mean.
**performance data has been associated with improvements. no substantial worries about unintended consequences.
**Increased use of screening since the measure has been implemented in all regions of Massachusetts from less than 5% to more than 70% demonstrate the utility of the tool and the measure and suggest national implementation would be useful.
**i think this is a useful measure and useful tool overall. the burden is relatively low, and it opens the door for open dialogue with the patient/family about various psychosocial issues. potential harms may include potential unnecessary referrals or over-utilzation of meds (false positives). my other concern is the PSC doesn't include questions surrounding suicide, self injurious behaviors, or substances.
**Potential for false positives, but outweighed by the benefit of early identification of issues.
**the unintended consequence is that more children are given ineffective and low quality treatment for behavioral health disorders which could result in iatrogenic complications as well as wasted time and money.
**I think future research should focus on a related potential measure about best practices that impact how many patients actually engage in and improve from mental health treatment over time. I hypothesize that HOW refills are made qualitatively and quantitatively impact this.
**Agree with usability.
**The only potential harm is if parents of youth who screen positive are not given appropriate and timely referrals. Physicians must review screening results same day it is completed (or day of well visit in the event the screen is taken in advance). The benefit of screening outweighs the harm of not screening or of gaps they may arise for those screened who don't get immediate referrals or feedback.
**The use of the measure could provide early detection and referral of mental health issues in children ages 3-17.9. Older children are often not screened and use of this measure would raise awareness to the provider about providing a screening.
**Do have issues with Usability. If I read this correctly would require manually reviewing records as there is no CPT codes specifically for this Screen. Developer should go through the process of getting a CPT codes for this so it could be easily extracted from claims.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0712 : Depression Utilization of the PHQ-9 Tool

Harmonization

Developer response to harmonization in current submission: The PSC screens for a broader band of problems (anxiety, behavior, etc.) and has a larger age range (3-17) than PHQ-9 (age range 12 to 17 and diagnosis of depression only). Studies show that the PSC identifies about 80% of youth with depression who are found with the PHQ-9, only about half of the youth with serious psychosocial problems on the PSC are identified with the PHQ-9 (Richardson et al., 2010).

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January 10, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

Evidence_Form_PSC_Resubmitted_20171214.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

1a Evidence (subcriterion 1a)

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

Measure Title: **Psychosocial Screening Using the Pediatric Symptom Checklist-Tool (PSC-Tool)**

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

Date of Submission: 11/8/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- 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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: Completion of the Pediatric Symptom Checklist Tool (PSC) by parents/guardians of youth ages 3.00 to 17.99 years

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

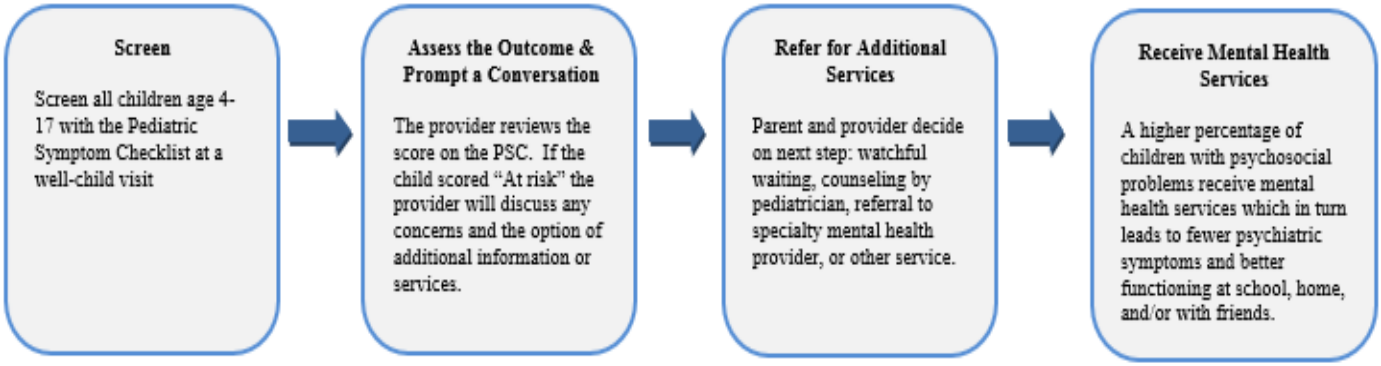


Figure 1 shows the logic model for the completion of the PSC tool as a required component of a pediatric WCV. This process measure increases the likelihood that psychosocial issues have been identified, explored, and dealt with during the WCV. There is now strong evidence that routine screening with the PSC Tool leads to a significant increase in the number of children with problems who receive outpatient mental health treatment. There is also moderate to strong evidence that children who are screened and receive services in this way show significant reductions in symptom and improved short- and long-term mental health and functional outcomes.

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	
Estimates of benefit and consistency across studies	
What harms were identified?	
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	

1a.4 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.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

The Pediatric Symptom Checklist (PSC) is a brief tool designed to be completed by parents of youth ages 4-17 years in conjunction with routine pediatric well child visits (WCV) to increase the chance that concerns parents may have about a child's psychosocial functioning are brought up during the WCV. Although the feasibility (Pourat, Zima, Marti, & Lee, 2017) and acceptability (Semansky, Koyanagi, & Vandivort-Warren, 2003) of the PSC as a clinical and research measure with diverse populations has been demonstrated in more than 180 studies over the past 30 years, (Haile et al., 2017) it is only in the past few years that evidence of the positive impact of routine psychosocial screening on child mental health outcomes has become available. Routine screening in pediatrics with the PSC and similar tools has been shown to be feasible on a statewide scale and sustainable over nearly a decade (Hacker et al., 2016; Savageau et al., 2016).

Screening has been shown to increase parent-provider communication about mental health concerns (Berger-Jenkins, McCord, Gallagher, & Olfson, 2012; Fortney et al., 2015; Hayutin, Reed-Knight, Blount, Lewis, & McCormick, 2009) as well as to identify at risk youth who are not in treatment but may benefit from mental health services (Hacker et al., 2014a; Hacker et al., 2014b). By generating a conversation with parents regarding their concerns about their children's psychosocial problems based on a risk score on the PSC, there is now strong evidence that children who

screen positive are more likely to be referred to and to receive mental health services (Hacker et al., 2014a; Hacker et al., 2014b)_ENREF_7.

Experimental studies have shown that patients in outpatient pediatrics who are screened with the PSC and have a range of psychosocial problems can be treated for these problems in the pediatric setting and that those who are treated show significantly better mental health outcomes than children who are randomized to usual care (Kolko et al., 2014; Kolko, Cheng, Campo, & Kelleher, 2011). These findings demonstrate the potential to bridge the gap between detecting psychosocial problems in pediatrics, increased mental health treatment, and improved psychosocial functioning in the immediate (Kolko et al., 2014; Kolko et al., 2011) and longer term (Kellam et al., 2014).

Most of the evidence for improved outcomes with screening at a WCV comes from a statewide program (Children's Behavioral Health Initiative) that began in Massachusetts in 2008, which mandates that children with Medicaid health insurance receive a brief formal validated screening measure as a part of the WCV. To date, more than 3 million screens have been completed in conjunction with more than 4 million WCV (http://www.mass.gov/eohhs/consumer/insurance/cbhi/) as a part of CBHI and the results of a related study suggest that the PSC is by far the most commonly used screening measure for children ages 3-17 in CBHI (Savageau, Simons, Lucke, Jellinek, & Murphy, 2017, May). Specifically, in a subsample of 3-17 year old children from that sample who received a formal screen at their WCV, two thirds (67%) of parents completed the PSC. In the discussion that follows, we use CBHI data for 3-17 year old children as a proxy for patterns that would be found in a PSC only sample.

Using administrative claims data from the first several years of the program based on cohorts of more than 100,000 visits, two studies documented a significant increase in routine screening and in the number of children receiving outpatient mental health services (Hacker et al., 2014a; Hacker et al., 2014b). These studies showed that more than 40% of the patients who screened positive had not been in treatment, (Hacker et al., 2014a) and that about 30% of the newly identified children went on to receive behavioral health services (Hacker et al., 2014b). A third study used a quasi-experimental design and a sample of almost 10 million WCV to compare pediatric screening and mental health service outcomes in two states over the two years before and after the mandatory screening program started in Massachusetts (Hacker et al., 2016). This study found that in Massachusetts the rate of outpatient behavioral health service utilization increased (from about 35 per 1,000 youths per month prior to the start of the mandate to about 50 per 1,000 youths per month after it). In contrast, however, in California, behavioral health service use remained at the same rate over the same four year period (Hacker et al., 2016). It may be important to note that the differential increase was in outpatient therapy only and not in emergency department, inpatient, or psychopharmacology treatment. In a state-funded evaluation of the CBHI program (Savageau et al., 2016), researchers from the University of Massachusetts Medical School confirmed the substantial and significant increases in formal psychosocial screening using a chart review of random samples of 2000 cases for the year before and two years after the start of CBHI. In a secondary analysis of the same data, these researchers assessed racial, ethnic, and language disparities in screening and services and found none (Savageau et al., 2017, May).

Although neither of these two samples permitted an assessment of differences between clinics in the rates of screening or outpatient mental health services, data posted on the CBHI website show clear and consistent differences in the state's six geographical regions in rates of screening, differences that have persisted over the course of the entire program (http://www.mass.gov/eohhs/consumer/insurance/cbhi/). The UMass researchers reported comparable and statistically significant differences in rates of screening by region in their chart review sample (Savageau et al., 2016).

To provide preliminary data on whether the kinds of differences documented on the state and regional level would be found at the clinic level, our research group at Massachusetts General Hospital obtained a sample of claims data for four MGH affiliated outpatient pediatric practices. We examined a sample of 10,827 youth ages 4-17 with MassHealth as their primary insurance and therefore required to receive behavioral health screening at their WCV. We focused on the sample of youth who had a WCV in each year (2014, 2015, 2016) and reported on the analyses from 2016 since this was the year that the hospital's new EMR (Epic) and its billing module were thought to be fully operational. As shown in Table 7, the results demonstrated that there were significant differences between clinics on the prevalence of screening at the WCV in 2016 (0% to 89.6%).

Table 7. Rates of screening at WCV in 2016, for children ages 4-17 in 4 clinics

	MH Screen at WCV in 2016	
Clinic		
A	3422 (86.1%)	
B	1215 (88.1%)	
C	106 (24.1%)	
D	0 (0.0%)***	
Total	4743 (69.7%)	

Note. N= children ages 4-17 with MASSHEALTH or MASSHEALTH PCC or NHP MASSHEALTH as primary insurance at each visit in 2014, 2015, and 2016
****p*<.001 $^{1}x^{2}$= 664.38 p<.001

These recent empirical studies supporting the measure rest on a large body of prior work in the form of recommendations by professional and advocacy organizations, requirements by insurers, reviews leading to policy changes by state governments, and related systematic reviews.

American Academy of Pediatrics Recommendations:

Although it has not provided a set of official guidelines, the American Academy of Pediatrics has endorsed routine psychosocial screening as a part of WCVs for children of all ages (Foy, Kelleher, Laraque, & Health, 2010) recently and for more than two decades, originally through its 1994 publication of Bright Futures (Green, 1994) which featured the PSC prominently, more recently through recommendations published in Pediatrics (Weitzman & Wegner, 2015), and through a toolkit on a CD designed to help practices do more to address psychosocial issues (AAP, 2015).

In 2010, the American Academy of Pediatrics (Foy et al., 2010) recommended that, "Pediatricians use validated instruments to screen all school aged children (5 years through adolescence) for symptoms of mental illness and impaired psychosocial functioning at health maintenance visits..." [S103]. The report based its recommendation on a review of the literature that demonstrated in numerous studies of the feasibility of using brief, psychometrically sound mental health screening tools in primary care settings. The report states that the literature "supports the conclusion that screening with a validated tool is useful in identifying children with mental health problems in a variety of settings. These settings include regular health maintenance visits" (Appendix S4 p.133). Most critically, this recommendation was based on "a number of studies [which] indicate that the use of screening methods improves identification of children in need of services" (Appendix S4 S134) and that "prevention and early intervention efforts targeted to children, youth, and families have been shown to be cost-effective, reducing use of more costly services such as welfare dependency and juvenile detention. Emotional and behavioral problems in young children may persist or worsen and adversely affect early and later school performance, and children from poor families are generally at greater risk. These findings suggest that early detection and intervention, particularly in low-income populations, may prevent or ameliorate mental health problems in children and adolescents" (Appendix S4; S135) (Foy et al., 2010).

EPSDT Regulations

Routine mental health screening as a part of WCV is a requirement of the US Medicaid program which provides health insurance coverage to 27 million children under age 18 in the United States. The guidelines governing Medicaid are known as Early and Periodic Screening, Diagnostic and Treatment (EPSDT) (Mann, 2013).

> The EPSDT program assures that health problems, including mental health and substance use issues, are diagnosed and treated early before they become more complex and their treatment more costly. Under the EPSDT benefit, eligible individuals must be provided periodic screening (well child exams) as defined by statute. One required element of this screening is a comprehensive health and developmental history including assessment of physical and mental health development. Part of this assessment is an age appropriate mental health and substance

> use health screening. As noted in the section above, early detection of mental health and substance use issues is important in the overall health of a child and may reduce or eliminate the effects of a condition if diagnosed and treated early. If, during a routine periodic screening, a provider determines that there may be a need for further assessment, an individual should be furnished additional diagnostic and/or treatment service (pg. 2).

Although mental health screening has been an explicit requirement for all WCV covered by Medicaid in all fifty states for nearly 30 years, compliance with this requirement had been uniformly low (Semansky et al., 2003) until a lawsuit in Massachusetts led to a consent decree in that state that mandated the use of a formal mental health screen in every WCV for every child covered by Medicaid. This program, known as the Children's Behavioral Health Initiative (Kuhlthau et al., 2011) also included a major increase in the accessibility and diversity of mental health services provided by the state. With approximately 3 million screens collected over the first 9 years of the CBHI program and more than half a dozen academic papers (Hacker et al., 2014a; Hacker et al., 2014b; Kuhlthau et al., 2011; Romano-Clarke et al., 2014; Savageau et al., 2016), the evaluations of CBHI have provided the strongest evidence to date on the need for and impact of routine psychosocial screening with the PSC and similar measures in pediatrics.

More support for routine psychosocial screening in pediatrics comes from Blue Ribbon commissions and advocacy groups. Mental Health America, the country's oldest mental health advocacy organization has recently made a major commitment to mental health screening (http://www.mentalhealthamerica.net/about-mha-screening). As of this writing more than 1.5 million screens have been collected as a part of it "B4Stage4" program which makes the PSC and several other measures available online for free.(J. M. Murphy et al., 2017) Over the past few months MHA has taken an even more active role, presenting a powerful position paper (http://www.mentalhealthamerica.net/positions/early-identification) in favor of early identification of mental health issues in young people to the new Interagency Serious Mental Illness Coordinating Committee (ISMICC) that was established by the 21st Century Cures Act. MHA has also been partnering with other mental health and substance use advocacy organizations to ask for inclusion of measures of screening rates and treatment outcomes in the CMS-AHIP Core Quality Measures Collaborative, arguing that the used of these measures should be tied to value-based payments for all health systems and providers to incentivize detection at the earliest point and focus on recovery as the goal http://www.mentalhealthamerica.net/blog/here's-what-we're-telling-new-ismicc-it-must-do

The president's New Freedom Commission on Mental Health (Hogan, 2003) also strongly recommended routine mental health screening as did the Institute of Medicine's influential literature review on "Preventing mental, emotional, and behavioral disorders among young people" (O'Connell, Boat, & Warner, 2009).

The Kennedy Forum (https://thekennedyforum.org/about/) has recently come out strongly in favor of routine psychosocial screening in pediatrics. In its Issue Brief titled "Fixing Behavioral Health Care in America: A National Call for Measurement-Based Care in the Delivery of Behavioral Health Services" the Kennedy Forum strongly endorsed the following policy: (Fortney et al.)

> All primary care and behavioral health providers treating mental health and substance use disorders should implement a system of measurement-based care whereby validated symptom rating scales are completed by patients and reviewed by clinicians during encounters. Measurement-based care will help providers determine whether the treatment is working and facilitate treatment adjustments, consultations, or referrals for higher intensity services when patients are not improving as expected. (p.7)

The Kennedy Forum based their recommendation on the literature that demonstrated the effectiveness, feasibility, and acceptability of Measurement-Based Care (MBC). In terms of effectiveness: "Virtually all randomized controlled trials with frequent and timely feedback of patient reported symptoms to the provider during the clinical encounter significantly improved outcomes or trended towards significance" (p.6) (Anker, Duncan, & Sparks, 2009; Bickman, Kelley, Breda, de Andrade, & Riemer, 2011; Brodey et al., 2005; Harmon et al., 2007; Hawkins, Lambert, Vermeersch, Slade, & Tuttle, 2004; Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Krägeloh, Czuba, Billington, Kersten, & Siegert, 2015; Lambert et al., 2002; K. P. Murphy, Rashleigh, & Timulak, 2012; Reese, Norsworthy, & Rowlands, 2009; Reese, Toland, Slone, & Norsworthy, 2010; Simon, Lambert, Harris, Busath, & Vazquez, 2012; Slade, Lambert, Harmon, Smart, & Bailey, 2008; Whipple et al., 2003). In terms of feasibility, "Measurement-based care can be

incorporated into routine care regardless of the characteristics of the patient population, or the treatment philosophy and training background of providers" (p.6) (Katzelnick et al., 2011; Sachs et al., 2003; Trivedi et al., 2006). In terms of acceptability, "MBC [Measurement-Based Care] has a high acceptance rate among patients and providers" (p.6) (Dowrick et al., 2009; Goldstein et al., 2011; Zimmerman & McGlinchey, 2008).

There have been two systematic reviews of routine psychosocial screening in pediatrics but since neither of them dealt with routine psychosocial screening as a process measure, they have been included in this "Other evidence" section of our proposal.

The first review, "Universal mental health screening in pediatric primary care: a systematic review" (Wissow et al., 2013) focused on the processes behind screening, how parents and youth are engaged, and how providers evaluate and use screening results and noted:

> The strongest conclusion that can be drawn from this review is that the existing literature on pediatric mental health screening processes for patient engagement and provider use is very limited. Key issues such as how to present screeners in ways that are not potentially damaging to therapeutic relationships (intrusive, culturally inappropriate, not confidential, etc.), or how to help providers make valid use of screening results, have not received systematic study (p.1145).

In terms of the focus of this proposal, the review did conclude that screening done with the Pediatric Symptom Checklist has been found to be associated with increased behavioral discussions between the pediatrician and the family (1147.e2) (1147.e6), an increase in chart rates for mental health concerns (1147.e2), and referrals for mental health care (1147.e5).

In the second systematic review, Lavigne and his associates (Lavigne, Meyers, & Feldman, 2016) explored the classification accuracy of the PSC and several other widely used measures in integrated primary care settings. Considering the choice of different criterion measures and different cutoff scores, the study concluded that although three measures (the CBCL, PSC, and SDQ) had achieved relatively high SE and SP values (.70), in one third to one half of the studies reviewed these instruments had not done so using the same cut off score simultaneously. The authors acknowledge that "the choice of a screening measure is not solely a matter of classification accuracy [and that a] number of other clinical and practical considerations will play an important role in developing a screening process that is useful and can be maintained over time." The authors quote the AAP recommendations (Foy et al., 2010) in considering more than a dozen contextual and technical factors that they consider just as important as screening accuracy.

The way this much larger array of factors can be considered in selecting a pediatric behavioral health screen is illustrated in the last review document to be discussed in this section, which concluded by selecting the PSC as the sole brief measure for a new statewide program that mandates the use of standardized outcome measures for all children receiving state funded mental health services that is just getting under way. The UCLA Center for Health Policy Research was charged by the California Department of Health Care Services (DHCS) with examining available tools for the measurement of mental health functioning for children and adolescents served by California's publicly funded pediatric and specialty mental health systems. UCLA used qualitative methods like focus groups and surveys of providers and consumers as well as scientifically rigorous methods like Delphi panels and systematic literature reviews to develop a recommendation for a tool that would fit the needs of multiple stakeholders in their state. The result was a 55 page evidence review (Pourat et al., 2017) that came to the conclusion that the PSC should be the sole brief pediatric behavioral health measure used in a statewide program of screening and outcomes evaluation. The Child and Adolescent Needs and Strengths (CANS) assessment was chosen the basis for a more in-depth rating. Although the program will begin by using the PSC only as a pre/post measure for children receiving publicly funded mental health services, the hope is that in future years other child serving programs like pediatrics, foster care, juvenile justice, etc. will also begin using the PSC routinely, thereby facilitating comparisons and clinical handoffs between these disparate agencies.

> "The panel received the available scientific evidence on all candidate tools and rated each tool individually based on four domains and on overall utility. The domains were: 1) effectiveness of care (face validity); 2) scientific acceptability; 3) usability; and 4) feasibility." The conclusion was that the "PSC-35 was consistently rated in the "very high" (average rating = 9) to "high" (average ratings = 7) ranges for overall utility, effectiveness of care,

scientific acceptability, usability and feasibility" (p. 31). The PSC-35 (parent version) was the only tool that satisfied all … minimum criteria for monitoring the effectiveness of publicly-funded child mental health care" (pg. 8).

"The PSC-35 is available for all age groups subject to the legislative mandate for outcome measurement, particularly very young children not covered in other tools examined. The PSC-35's focus on current (rather than past or retrospective) child mental health status is an important consideration because a child's episode of care varies in length and there may be gaps in care due to various barriers in access or other issues. The PSC-35 was the only high-scoring tool that had the capacity to measure clinical outcomes at chronologic time points that could potentially align with the receipt of recommended care or adherence to quality indicators within a child's unique episode of care. This capacity is important because conclusions about whether or not care is effective require that changes in clinical outcomes be interpreted within the context of the quality of care delivered" (pg. 38).

Based on these recommendations, the California state legislature voted $15,000,000 for fiscal 2018 to fund the training and development of a data collection infrastructure that will be the basis of a new statewide child mental health evaluation program that may eventually encompass all of child serving agencies in the state.

1a.4.2 What process was used to identify the evidence?

Literature synthesis on the strength and benefits of psychosocial screening.

1a.4.3. Provide the citation(s) for the evidence.

AAP. (2015). Addressing Mental Health Concerns in Primary Care: A Clinician's Toolkit.

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of consulting and clinical psychology, 77*(4), 693.

Berger-Jenkins, E., McCord, M., Gallagher, T., & Olfson, M. (2012). Effect of routine mental health screening in a low-resource pediatric primary care population. *Clinical pediatrics, 51*(4), 359-365.

Bickman, L., Kelley, S. D., Breda, C., de Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services, 62*(12), 1423-1429.

Brodey, B. B., Cuffel, B., McCulloch, J., Tani, S., Maruish, M., Brodey, I., & Unützer, J. (2005). The acceptability and effectiveness of patient-reported assessments and feedback in a managed behavioral healthcare setting. *The American journal of managed care, 11*(12), 774-780.

Dowrick, C., Leydon, G. M., McBride, A., Howe, A., Burgess, H., Clarke, P., . . . Kendrick, T. (2009). Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *Bmj, 338*, b663.

Fortney, J., Sladek, R., Unützer, J., Kennedy, P., Harbin, H., Emmet, B., . . . Carneal, G. Fixing Behavioral Health Care in America A National Call for Measurement-Based Care in the Delivery of Behavioral Health Services. Issue Brief Released by The Kennedy Forum.

Fortney, J., Sladek, R., Unützer, J., Kennedy, P., Harbin, H., Emmet, B., . . . Carneal, G. (2015). Fixing Behavioral Health Care in America: A National Call for Measurement-Based Care in the Delivery of Behavioral Health Service. *The Kennedy Forum*.

Foy, J. M., Kelleher, K. J., Laraque, D., & Health, A. A. o. P. T. F. o. M. (2010). Enhancing pediatric mental health care: strategies for preparing a primary care practice. *Pediatrics, 125*(Supplement 3), S87-S108.

Goldstein, L. A., Gibbons, M. B. C., Thompson, S. M., Scott, K., Heintz, L., Green, P., . . . Crits-Christoph, P. (2011). Outcome assessment via handheld computer in community mental health: Consumer satisfaction and reliability. *The journal of behavioral health services & research, 38*(3), 414-423.

Green, M. (1994). *Bright Futures: Guidelines for Health Supervision of Infants, Children, and Adolescents*: ERIC.

Hacker, K. A., Penfold, R., Arsenault, L., Zhang, F., Murphy, M., & Wissow, L. (2014a). Screening for behavioral health issues in children enrolled in Massachusetts Medicaid. *Pediatrics, 133*(1), 46-54.

Hacker, K. A., Penfold, R., Arsenault, L. N., Zhang, F., Soumerai, S. B., & Wissow, L. S. (2016). The Impact of the Massachusetts Behavioral Health Child Screening Policy on Service Utilization. *Psychiatric Services, 68*(1), 25-32.

Hacker, K. A., Penfold, R. B., Arsenault, L. N., Zhang, F., Murphy, M., & Wissow, L. S. (2014b). Behavioral health services following implementation of screening in Massachusetts Medicaid children. *Pediatrics*, peds. 2014-0453.

Haile, H., Lucke, C., Abel, M., McCarthy, A., Chiang, C., Kamin, H., & Murphy, J. (2017). A Review of Recent Research on the Pediatric Symptom Checklist (PSC) 2001-2017. *Unpublished manuscript*.

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist–client feedback and clinical support tools. *Psychotherapy research, 17*(4), 379-392.

Hawkins, E. J., Lambert, M. J., Vermeersch, D. A., Slade, K. L., & Tuttle, K. C. (2004). The therapeutic effects of providing patient progress information to therapists and patients. *Psychotherapy research, 14*(3), 308-327.

Hayutin, L. G., Reed-Knight, B., Blount, R. L., Lewis, J., & McCormick, M. L. (2009). Increasing parent–pediatrician communication about children's psychosocial problems. *Journal of Pediatric Psychology, 34*(10), 1155-1164.

Hogan, M. (2003). The President's New Freedom Commission: recommendations to transform mental health care in America. . *Psychiatric Services 54*, 1467-1474.

Katzelnick, D. J., Duffy, F. F., Chung, H., Regier, D. A., Rae, D. S., & Trivedi, M. H. (2011). Depression outcomes in psychiatric clinical practice: using a self-rated measure of depression severity. *Psychiatric Services, 62*(8), 929-935.

Kellam, S. G., Wang, W., Mackenzie, A. C., Brown, C. H., Ompad, D. C., Or, F., . . . Windham, A. (2014). The impact of the Good Behavior Game, a universal classroom-based preventive intervention in first and second grades, on high-risk sexual behaviors and drug abuse and dependence disorders into young adulthood. *Prevention Science, 15*(1), 6-18.

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *The British Journal of Psychiatry, 195*(1), 15-22.

Kolko, D. J., Campo, J., Kilbourne, A. M., Hart, J., Sakolsky, D., & Wisniewski, S. (2014). Collaborative care outcomes for pediatric behavioral health problems: a cluster randomized trial. *Pediatrics*, peds. 2013-2516.

Kolko, D. J., Cheng, Y., Campo, J. V., & Kelleher, K. (2011). Moderators and predictors of clinical outcome in a randomized trial for behavior problems in pediatric primary care. *Journal of Pediatric Psychology, 36*(7), 753-765.

Krägeloh, C. U., Czuba, K. J., Billington, D. R., Kersten, P., & Siegert, R. J. (2015). Using feedback from patient-reported outcome measures in mental health services: a scoping study and typology. *Psychiatric Services, 66*(3), 224-241.

Kuhlthau, K., Jellinek, M., White, G., VanCleave, J., Simons, J., & Murphy, M. (2011). Increases in behavioral health screening in pediatric care for Massachusetts Medicaid patients. *Archives of pediatrics & adolescent medicine, 165*(7), 660-664.

Lambert, M. J., Whipple, J. L., Vermeersch, D. A., Smart, D. W., Hawkins, E. J., Nielsen, S. L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical psychology & psychotherapy, 9*(2), 91-103.

Lavigne, J. V., Meyers, K. M., & Feldman, M. (2016). Systematic Review: Classification Accuracy of Behavioral Screening Measures for Use in Integrated Primary Care Settings. *Journal of Pediatric Psychology, 41*(10), 1091-1109. doi: 10.1093/jpepsy/jsw049

Mann, C. (2013). CMCS Informational Bulletin; Prevention and Early Identification of Mental Health and Substance Use Conditions.

Murphy, J. M., Nguyen, T., Lucke, C., Chiang, C., Plasencia, N., & Jellinek, M. (2017). Adolescent self-screening for mental health problems; Demonstration of an internet-based approach. *Academic Pediatrics*.

Murphy, K. P., Rashleigh, C. M., & Timulak, L. (2012). The relationship between progress feedback and therapeutic outcome in student counselling: A randomised control trial. *Counselling Psychology Quarterly, 25*(1), 1-18.

O'Connell, Boat, & Warner. (2009). Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities. Washington, D.C.: Committee on the Prevention of Mental Disorders and Substance Abuse Among Children Youth and Young Adults.

Pourat, N., Zima, B., Marti, A., & Lee, C. (2017). California Child Mental Health Performance Outcomes System: Recommendation Report. *UCLA Center for Health Policy Research*.

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy: Theory, research, practice, training, 46*(4), 418.

Reese, R. J., Toland, M. D., Slone, N. C., & Norsworthy, L. A. (2010). Effect of client feedback on couple psychotherapy outcomes. *Psychotherapy: Theory, research, practice, training, 47*(4), 616.

Romano-Clarke, G., Tang, M. H., Xerras, D. C., Egan, H. S., Pasinski, R. C., Kamin, H. S., . . . Murphy, J. M. (2014). Have rates of behavioral health assessment and treatment increased for Massachusetts children since the Rosie D. decision? A report from two primary care practices. *Clinical pediatrics, 53*(3), 243-249.

Sachs, G. S., Thase, M. E., Otto, M. W., Bauer, M., Miklowitz, D., Wisniewski, S. R., . . . Frank, E. (2003). Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biological psychiatry, 53*(11), 1028-1042.

Savageau, J. A., Keller, D., Willis, G., Muhr, K., Aweh, G., Simons, J., & Sherwood, E. (2016). Behavioral Health Screening among Massachusetts Children Receiving Medicaid. *The Journal of Pediatrics, 178*, 261-267.

Savageau, J. A., Simons, J., Lucke, C., Jellinek, M., & Murphy, J. M. (2017, May). Assessing Disparities in Mental Health Screening and Services Before and After Implementation of an Innovative Statewide Program for Children with Medicaid. *Accepted for presentation at Pediatric Academic Societies (PAS) 2017 Meeting, San Francisco, CA.*

Semansky, R. M., Koyanagi, C., & Vandivort-Warren, R. (2003). Behavioral health screening policies in Medicaid programs nationwide. *Psychiatric Services, 54*(5), 736-739.

Simon, W., Lambert, M. J., Harris, M. W., Busath, G., & Vazquez, A. (2012). Providing patient progress information and clinical support tools to therapists: Effects on patients at risk of treatment failure. *Psychotherapy research, 22*(6), 638-647.

Slade, K., Lambert, M. J., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate electronic feedback and revised clinical support tools. *Clinical psychology & psychotherapy, 15*(5), 287-303.

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., . . . McGrath, P. J. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR* D: implications for clinical practice. *American journal of Psychiatry, 163*(1), 28-40.

Weitzman, C., & Wegner, L. (2015). Promoting optimal development: Screening for behavioral and emotional problems. *Pediatrics*, peds. 2014-3716.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*(1), 59.

Wissow, L., Brown, J., Fothergill, K., Gadomski, A., Hacker, K., Salmon, P., & Zelkowitz, R. (2013). Universal Mental Health Screening in Pediatric Primary Care: A Systematic Review. *Journal of the American Academy of Child & Adolescent Psychiatry, 52*(11), 1134 - 1147.e1123.

Zimmerman, M., & McGlinchey, J. B. (2008). Depressed patients' acceptability of the use of self-administered scales to measure outcome in clinical practice. *Annals of Clinical Psychiatry, 20*(3), 125-129.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

Psychosocial problems in children are common and treatable with prevalence estimates of about 12% of all children and adolescents (Gardner, Lucas, Kolko, & Campo, 2007; Kelleher et al., 1997; Murphy et al., 2016). Studies have shown that children with these problems are often unrecognized by their pediatricians (~50% of cases) (Kelleher et al., 1997) and that only a fraction of them receive treatment (Hacker et al., 2014b; Kelleher et al., 1997). Children with psychosocial problems are more likely to have poorer health, academic, behavioral, and social outcomes in both the short and long term (Murphy et al., 2015). Children who receive psychosocial screening as a part of pediatric well child visits are more likely to receive outpatient mental health services (Hacker et al., 2014a; Hacker et al., 2014b; Savageau et al., 2016) than are children who are not screened. As the dates of the studies just cited attest, it is only within the last three years that strong evidence documenting the relationship between psychosocial screening and increased mental health treatment has become available.

A series of RCT studies by Kolko and his associates have shown that pediatric outpatients with a wide range of problems who are found to be at risk when screened with the PSC and go on to receive pediatric office based mental health interventions have significantly lower mental health symptom scores and better functioning at immediate and longer term follow up than do similar outpatients randomized to treatment as usual (Kolko et al., 2014; Kolko, Campo, Kelleher, & Cheng, 2010). For these reasons, we believe that an increase in mental health treatment is the most appropriate (and a measurable) benchmark for assessing the positive impact of routine psychosocial screening. The logic model for screening in pediatrics is that more children will receive help, fewer children will develop mental, emotional, and behavioral disorders (Guzmán et al., 2015; Kieling et al., 2011), and more children who received help will enjoy better life outcomes (Kellam et al., 2014).

Requiring screening for psychosocial problems as part of routine well child care in pediatrics is one of the most frequently recommended ways to improve recognition and intervention for such problems (Hacker et al., 2014a) and an increasing number of states (Massachusetts (Savageau et al., 2016)), insurers (Medicaid/EPSDT (Mann, 2013)), standard setting organizations (American Academy of Pediatrics (Foy, Kelleher, Laraque, & Health, 2010; Weitzman & Wegner, 2015)), blue ribbon panels (President's New Freedom Commission on Mental Health (Hogan, 2003) (Institute of Medicine (O´Connell, Boat, & Warner, 2009)), and advocacy organizations such as the Kennedy Forum (Fortney et al., 2015) and Mental Health America (http://www.mentalhealthamerica.net/positions/early-identification) have now required, endorsed, or recommended the principle of including a psychosocial screen as a part of every well child visit for children aged 3-17.

The PSC is probably the most frequently recommended and widely used tool for routine psychosocial screening in pediatrics (Semansky, Koyanagi, & Vandivort-Warren, 2003), with the Strengths and Difficulties Questionnaire (Goodman, Meltzer, & Bailey, 1998) and Child Behavior Checklist (Achenbach, 2009) instruments that are similar in many ways and also frequently mentioned and validated in this context. Many of the endorsements noted above include these three and/or a few other instruments.

The reference list is included in the attached appendix.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.

Several studies based on data from the Massachusetts Medicaid (MassHealth) pediatric behavioral health screening program and the Children's Behavioral Health Initiative (CBHI), demonstrate the currently wide variation in the rates of mental health screening with formal, standardized tools as well as the possibility of improvement and the potential benefits of doing so.

Data source 1:

Summary data from the CBHI Behavioral Health Screening Cumulative Quarterly Report; posted on BHSCQR website

Measurement Period dates of service from 1/1/2008 to 3/31/2017

Data Source 1a: Statewide data for all children

Denominator/Well child visits for all children .5 -20 years of age: 4,721,790

Numerator (screens with visit): 2,965,923

Statewide average: 62.8%

Minimum: 14.2%

Maximum: 71.9%

Standard Deviation: 12.4%

95% Confidence interval: 58.8% to 66.8%

The CBHI BHSCQR also presents the same statewide Medicaid screening data broken down for each of the state's regions for each of the 37 calendar quarters from Q1 (January through March of 2008) to Q37 (January through March of 2017) of CBHI. Table 1 (below) reports the rates of screening in the state's six regions with data from the four quarters in each year averaged for simpler presentation in this proposal. Because the BHSCQR website does not break out the data by age group, the information in the table below is for the full sample of all ages, but the trends should be large enough that that the patterns shown below should be quite similar to those for just the 3-17 year olds. As the table shows, the range in rates of screening vary widely, from a low of 21.55 in the Boston region in 2008 to a high of 85.27 in the Western region in first quarter of 2017. The percentages of well child visits with screens goes up substantially in all regions over the first few years of the initiative but after that the rank ordering remains relatively consistent across regions. Western Massachusetts always has the highest rate of screening, Metro West and Boston the lowest and Northeast, Southeast, and Central Massachusetts in the middle.

Table 1. Rates of Screening in different regions of Massachusetts from 2013-2017 by Region

Region	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017*
Western	38.14%	66.13%	75.34%	79.16%	79.79%	80.33%	81.62%	83.76%	81.27%	85.27%
Central	32.61%	51.99%	60.61%	63.22%	67.23%	70.59%	71.09%	70.90%	72.33%	72.31%
Northeast	32.70%	56.36%	60.93%	60.31%	55.35%	65.08%	67.16%	68.43%	71.46%	67.38%
Metro West	25.89%	47.45%	51.92%	51.80%	57.72%	60.47%	59.83%	54.90%	49.84%	47.72%
Southeast	36.88%	61.89%	70.57%	75.94%	78.18%	80.19%	80.66%	77.87%	75.91%	73.95%
Boston	21.55%	45.91%	55.28%	57.26%	60.53%	64.49%	67.57%	65.12%	63.27%	58.76%

Note. Rates of screening were calculated by the year from 2008-2016 and for the first quarter of 2017 *Data available only for the 1st quarter of 2017

Table 2 below presents the distribution of rates of screening in greater detail, taken from this time directly from BHSCQR website for all 222 measurement points (37 quarters in all six regions), illustrating even more dramatically the wide range of rates of screening across the state and over nearly a decade.

Table 2 The distribution of screening rates by decile broken down by quarters for each region from January 2008 through March 2017 (6 regions x 4 quarters x 9.25 years) for children of all ages

0-9%	1
10-19%	5
20-29%	4
30-39%	8
40-49%	23
50-59%	40
60-69%	63
70-79%	58
80-89%	20
90-100%	
Total (Regions*Quarters*Years)	222
Min:	8.21%
Max:	85.65%
Range:	77.44%
Median:	64.53%

Data Source 1b: Statewide data broken down by for children ages 3-17 years

Denominator /Well child visits for all children 3 -17 years of age: 2,361,475

Numerator (screens with visits): 1,681,764

Statewide average: 71.2%

Minimum: 39.98%

Maximum: 79.14%

Standard Deviation: 11.3%

95% Confidence Interval: 64.2% to 78.2%

We included data source 1b focusing just on the 3-17 year old children in the BHSCQR website data as this is the age group that was screened with the PSC, and prior studies (Hacker et al., 2016; Savageau et al., 2016) have shown that the PSC was the measure used for 67% of the children in the CBHI 3-17 year old age group. By multiplying the total number of screens in this age group by 67%, we can estimate that approximately 1,126,782 PSC's were administered over the first 9.25 years of CBHI.

Table 3 (below) shows the Massachusetts statewide number of well child visits, number of screens, and percent of visits with screens, for just the 3-17 year old (PSC screened) children with Medicaid from January 2008 (start of CBHI) to March of 2017. Data in this table are taken directly from the CBHI BHSCQR but with the four quarters of each year aggregated together so that the totals for each year could be seen more clearly. As the table shows, the rate of screening rose from approximately 39.98% for its first year (2008) to 65.72% for its second year to over 70% for its third year, and then remaining in the 70% range in all of the six years since. Not shown in this table but present in the data shown on the CBHI BHSCQR are the figures for the first quarter of 2008 (which show a base rate of 17.8% during the first three months of the program). Although not posted on the CBHI BHSCQR but reported by two different groups (Hacker et al., 2016; Savageau et al., 2016) with access to claims data for 2007 is that the rate of formal screening during the year prior to the start of CBHI was less than 5%. A rate of formal screening that started and then remained at less than

5% was also reported for the state of California (which had no requirement for the use of formal screens) for 2008 and 2009 when the rate in Massachusetts had climbed to about 65%.

Table 3. Statewide rates of formal psychosocial screening for 3-17 year olds during WCV from 1/1/2007 to 3/31/2017

Year

	Denominator (Total Visits)	Numerator (Total with Screens)	% Visits with Screens
12/31/07-12/31/2008	202,376	80,910	39.98%
1/1/2009-12/31/2009	219548	144276	65.72%
1/1/2010-12/31/2010	234823	171114	72.87%
1/1/2011-12/31/2011	243209	179305	73.72%
1/1/2012-12/31/2012	252357	190286	75.40%
1/1/2013-12/31/2013	265826	208549	78.45%
1/1/2014-12/31/2014	287690	228467	79.14%
1/1/2015-12/31/2015	290313	217282	74.84%
1/1/2016-12/31/2016	300307	215611	71.80%
1/1/2017-3/31/2017	65,026	45,964	70.69%
Total	2,361,475	1,681,764	71.22%

Data Source 2: CBHI Cohort Data for chart review sample

Measurement Period dates of service in 2007, 2010, and 2012

Denominator /Well child visits for all children: 4,977

Numerator (screen at WCV): ~1,700

Average: 51.7%

Min: 1.5%

Max: 88.9%

Standard Deviation : 35.1%

95% Confidence interval: 35.5% to 87.2%

This second data source is a ~ 6000 visits subsample of the CBHI statewide data retrieved from chart reviews supplemented by administrative claims data.

Differences in rates of screening in three cohorts of about 2000 cases of pediatric outpatients (age 4-17 years) seen before (2007) and after (2010, 2012) the start of CBHI. As shown in Table 4, consistent with the statewide CBHI data, the Southeast and Western regions consistently demonstrated significantly higher rates of screening, the Northeast and Metro West demonstrated the lowest rates of screening, and Central Massachusetts and Boston demonstrated screening rates in the middle.

Table 4. Rates of Formal Screening for each year for children ages 4-17

	Formal Screen in 2007		Formal Screen in 2010		Formal Screen in 2012	
Region	No	Yes *1	No	Yes ***2	No	Yes ***3
Western	139(95.2%)	7 (4.8%)	22(11.9%)	163(88.1%)	33(15.2%)	184(84.8%)
Central	64(94.1%)	4(5.9%)	30(23.3%)	99(76.7%)	31(23.9%)	99(76.1%)
Northeast	125 (91.2%)	12(8.8%)	63(37.5%)	105(62.5%)	71(39.0%)	111(61.0%)
Metro West	84 (94.4%)	5(5.6%)	54(32.5%)	112(67.5%)	62(37.4%)	104(62.7%)
Southeastern	197(98.5%)	3(1.5%)	52(23.0%)	174(77.0%)	27(11.1%)	217(88.9%)
Boston	149(98.0%)	3(2.0%)	51(25.4%)	150(74.6%)	33(18.2%)	148(81.8%)

2007: 1x2=13.33 p=.021 2010: 2x2=36.33 p<.001 2012: 3x2=75.21 p<.001

Data Source 3: Medicaid screening data from four Massachusetts General Hospital outpatient clinics

Measurement Period dates of service from 7/1/2014 to 12/31/2016

Denominator /Well child visits for all children 4-17 years of age: 10,334

Numerator (screen at WCV): 7,915

Average: 76.6%

Min: 9.4%

Max: 91.7%

Standard Deviation: 38.5%

95% Confidence interval: 38.8% to 100.0%

One of the most important criteria for a quality measure is differences in performance across sites. Although the data from the CBHI BHSCQR summarized above provide strong evidence that there are consistent differences in screening rates for different regions of the state and different age groups, our lack of access to the actual data made it impossible to explore differences between different pediatric practices, one of the intended uses of this type of quality measurement. Since it was not possible for us to obtain the actual data from the state of Massachusetts and since the sample from Data Source 2 was too small to permit analyses by clinic, we turned to our own hospital system and were able to obtain billing data for pediatric psychosocial screens for a relatively large sample of well child visits in four MGH-affiliated outpatient pediatric clinics. Using the same CPT billing code for screening (96110) used by the state and focusing only on children with Medicaid health insurance, we obtained data on 10,827 children aged 4-17 who had at least one pediatric well child visit from July 2014 through December 2016 in one of the four clinics.

Differences in rates by clinic: As shown in Table 5 below, the differences in rates of mental health screening at well-child visits were substantial and statistically significant between these four clinics. In 2016, for example, rates ranged from 0% to 88.1% ($p < .001$) to in the four clinics. Over the three years of data, the rank orders of screening rates among the four clinics were relatively constant with Clinic A and B as always the highest and C the lowest in each of the three years (not shown).

2016 (n=6,801 children): January 1 2016 – December 31 2016

Variable	Overall*	Clinic A	Clinic B	Clinic C	Clinic D
Billed for MH screen (yes)	69.7% (4743)	86.1% (3422)	88.1% (1215)	0.0% (0)	24.1% (106)

p<.001

The reference list is included in the attached appendix.

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Differences by age:

As presented below in Table 6 (below) from Data Source 1, the CBHI BHSCQR shows significant differences in rates of screening by age group with very young (< 3) children and older (> 17) youth less likely to be screened than 3-17 year olds. Even within the 3-17 year old group there were differences in screening rates with younger (3-6 year olds) and older (13 to 17 year olds) patients showing lower rates of screening than 7-12 year olds (72.41% vs 67.32% vs 73.08%) respectively.

Table 6 (December 31 2007 - March 31 2017)

Age	Total Visits	Total Visits with Screens	% Visits with Screens
<6mos to 2 years	2,136,135	1,205,607	56.44%
<6mos	882,43	336,179	38.10%
6mos to 2yrs	1,253,701	869,428	69.35%
3 yrsto 17 yrs	2,361,475	1,681,764	71.22%
3yrs to 6 yrs	776,570	562,311	72.41%
7 yrs to 12 yrs	911,693	666,266	73.08%
13 yrs to 17 yrs	673,212	453,187	67.32%
18 yrs to 20 yrs	224,180	78,552	35.04%
Total	4,721,790	2,965,923	62.80%

Lack of disparities by race, ethnicity, and language:

Although none of the published papers on the CBHI sample (Hacker et al., 2014a; Hacker et al., 2014b; Savageau et al., 2016) explore data on disparities by race, ethnicity, or language, the chart review study of a subsample of these cases from Data Source 2 (Savageau et al., 2016; Savageau, Simons, Lucke, Jellinek, & Murphy, 2017, May) explored screening by demographics in a subsample of ~ 6000 visits from 2007, 2007, and 2010. As shown in Table 7 below, there were no significant disparities by race, ethnicity, or language. It may be important to note that disparities by socioeconomic status cannot be meaningfully assessed in this sample since by CBHI is a program only for children with Medicaid and SES is confounded with insurance type.

Table 7. Lack of Significant Disparities of Children 4-17 years insured by MassHealth, FYs 2007*, 2010, and 2012

	Had a Formal Screen in 2007	Had a Formal Screen in 2010	Had a Formal Screen in 2012
Race			
White	7 (2.9%)	374 (73.6%)	397 (73.7%)
Non-White	20 (6.4%)	276 (77.7%)	290 (79.2%)
Ethnicity			
Non-Hispanic	17 (4.4%)	321 (75.3%)	369 (78.7%)
Hispanic	10 (5.8%)	224 (79.4%)	246 (80.9%)
Primary Language			
English	24 (3.8%)	676 (75.4%)	716 (76.8%)
Non-English	10 (6.5%)	128 (71.1%)	149 (76.4%)

The reference list is included in the attached appendix.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

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 sub criteria 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):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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.)*

http://www.massgeneral.org/psychiatry/services/psc_home.aspx

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)*

No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

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) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Number of patients with documentation that the PSC tool was administered as part of the well child visit.

S.5. 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, time period for data collection, 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 (S.14).

Depending on the system, patients passing this quality measure are identified either through a review of administrative claims or the medical record. In claims data, the presence of a CPT code for screening (96110 in Massachusetts and many other states) on the same day as the WCV is required. In a chart review, the presence of a PSC score or PDF scan of it in the progress note, or score shown in the visit template or flowsheet documents the completion of the screen on the same day of the WCV. To receive credit, progress notes must indicate the name of the specific measure and actual score (eg, PSC given, score = not at risk).

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

Number of patients aged 3.00 to 17.99 seen for a pediatric well-child visit.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Cases are identified from administrative data for site. Number of unique patients ages 3.00 to 17.99 seen for a well-child visit (CPT 99381-99394) in a defined evaluation period, often a year.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

No exclusions.

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

Step 1. Count number of children aged 3-17 seen for a well child visit in state, region, clinic or other group during defined period (often, one year) using administrative data (CPT 99381-99394). N=total population. This is the denominator.

Step 2. Assess whether PSC was administered as a part of WCV, for the eligible population, using patient claims data or chart for indicator status. Pass if documentation that screen was given on the day of the WCV is present.

Step 3. Compute numerator = count of patients with completed PSC.

Step 4. Calculate clinic or other entity rate as numerator/denominator. No risk adjustment.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

This measure and its denominator are not based on samples. This PRO measure is based on a parent or child completing the PSC (no proxy) and noting its presence/absence. Missing data (no administration of PSC) is managed by the inclusion of patients without a completed PSC in the denominator.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Claims, Electronic Health Records, Paper Medical Records

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

In administrative data:

If patient age => 3.0 & age =< 17.99; claim for well child visit (99382 or 99383 or 99385 or 99392 or 99393 or 99394), assess presence of CPT 96110 code for screening.

In medical record (paper or electronic):

If patient age => 3.0 & age =< 17.99; claim for well child visit (99382 or 99383 or 99385 or 99392 or 99393 or 99394), assess progress note, templated note, flowsheet, scanned in PSC, for evidence that screen was administered.

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No data collection instrument provided

S.20. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Clinician : Group/Practice, Facility, Population : Regional and State

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Outpatient Services

If other:

S.22. 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

2. Validity – See attached Measure Testing Submission Form

Measure_Form_PediatricSymptomChecklist.docx

2.1 For maintenance of endorsement

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 For maintenance of endorsement

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**: 3332**

Measure Title: **Psychosocial Screening Using the Pediatric Symptom Checklist-Tool (PSC-Tool)**

Date of Submission: 11/8/2017

Type of Measure:

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

Instructions

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

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

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

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

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

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for*

measure implementation. ***If different Data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

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

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

This measure is in full implementation in the Children's Behavioral Health Initiative (CBHI), a program that requires that all 6 month to 20.99 year old youth covered by Massachusetts Medicaid (MassHealth) be screened with one of a small number of validated and endorsed behavioral health (BH) measures as a part of pediatric well child visits (WCV). In order to meet all of the scientific acceptability testing requirements needed for NQF measure consideration, three sources of data on the CBHI BH screening data are discussed in this Measure Testing Form.

Data source 1: Summary data from CBHI pediatric behavioral health screening website

MassHealth posts a running summary of administrative claims data (Behavioral Health Screening Cumulative Quarterly Report; BHSCQR) for all well child visits for children with Medicaid in Massachusetts since the start of CBHI in January 2008 on a publicly accessible website (http://www.mass.gov/eohhs/docs/masshealth/cbhi/reports/bh-screening.pdf).

The CBHI BHSCQR summarizes the number of WCV's and the number and percentage of these visits with behavioral health screens and several other variables by quarter over the course of the initiative. Data are presented for the state as a whole and also broken down by age group and region of the state.

Data source 2: Chart review sample of 6000 cases from before and after start of CBHI

An external evaluation of CBHI was conducted by a team of researchers from the University of Massachusetts Medical School under a contract with Massachusetts state Medicaid (Savageau et al., 2016) using a combination of claims data and paper chart reviews by Gold Standard Testing certified RN chart abstractors. The study population for each period consisted of MassHealth-enrolled children at least 6 months old but under the age of 21 years. The final inclusion criterion for each period required children/adolescents to have a paid claim for a WCV, identified using current procedural terminology and *International Classification of Diseases, Ninth Revision, Clinical Modification* diagnosis codes. Based on age group stratification identified from American Academy of Pediatrics periodicities for WCVs and recommendations for MassHealth-approved standardized screening tools, stratified random sampling selected 500 members from each of 4 age groups (i.e., 6 months-2 years, 3-5 years, 6-11 years, and 12-20 years) in each study period, resulting in an initial total sample of 2000 members per year. The findings we present in this report are based on a secondary analysis of the same data conducted for us by the same researchers from the University of Massachusetts Medical School who had conducted the original evaluation.

Data source 3: Claims data and chart reviews for WCV from four MGH outpatient pediatric clinics

In a data set of WCV over 3 years for ages 4-17 with Medicaid from four Massachusetts General Hospital-affiliated outpatient clinics, we analyzed WCV and BH screening billing claims data to assess differences in rates of screening at the individual case and clinic level

Although Data sources 1 and 2 had provided our group with information on many WCV and screening variables and permitted analysis by region of state and age group, we did not have access to data at the individual patient or clinic level. To obtain an initial estimate of whether it was possible to identify differential rates of BH screening at the clinic level, we were able to obtain billing data for WCV and BH screens for a sample of WCV for 10,334 children in four MGH-affiliated outpatient pediatric clinics from 7/1/2014-12/31/2016. Using the same CPT billing code for screening (96110) used in Data sources 1 and 2, and focusing only on children with Medicaid health insurance, we were able to perform our analyses on subsamples of these individual cases as described below. For three of these clinics we were able to review the EMR's of a subsample of the cases to assess the reliability and validity of the PSC as a process measure.

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

Data source 1: Medicaid claims data for all 6 month to 20 year old patients seen for well child visits in the state of Massachusetts from 1/1/2008-3/31/2017.

Data source 2: Random samples of 2000 visits per cohort drawn from Data Source 1 for 3 cohorts: 7/1/2006-6/30/2007; 7/1/2009-6/30/2010; 7/1/2011-6/30/2012.

Data source 3: Billing and demographic data extracted from Partners Healthcare System data warehouse for claims for WCV and BH screens and demographic data for all 4-17 year old patients from four MGH-affiliated outpatient pediatric clinics from 7/1/2014-12/31/2016.

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

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.20*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☒ other: Population: Regional and State	☒ other: Population: Regional and State

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

Data source	Entity Type	Size, Location and Level of Data Included
1. Pediatric WCV claims data posted on CBHI BHSCQR website	State Medicaid	Massachusetts state and 6 regions
2. Pediatric WCV and BH screening claims data and statewide chart audit of randomly selected subsamples of claims for WCV	State Medicaid	Massachusetts state and 6 regions
3. Partners Healthcare System WCV and BH claims data and chart reviews for a subsample of cases	Health system	Massachusetts General Hospital, four outpatient pediatric clinics

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

Data source 1: The base sample is claims data on 4,721,790 visits from the entire state of Massachusetts for whom WCV claims for youth 6 months to 20 years of age were submitted. These data include all clinics and practices in the state. Since these data were presented for all six regions of state for all age groups, in some analyses we reported on just the 2,361,475 WCV and 1,681,764 screens completed for 3-17 year olds, and in other analyses, just the estimated 1,582,1881 WCV and 1,126,782 screens that were based on the PSC rather than other instruments.

Data source 2: A 6,000-case sample was drawn from Data source 1, with stratified random samples of 2000 cases drawn for each cohort for the FY 2007, 2010, and 2012 cohorts, with 500 cases each for 4 age groups (6 months to 2 years, 3 to 5 years; 6-11 years, and 12-20 years). Demographic data from this sample are presented below:

Table 1. Sociodemographic characteristics of children 4-17 years insured by MassHealth, FYs 2007, 2010, and 2012 (from Savageau, et al, 2016).

Characteristics	FY2007 (N=1336) N (%)	FY2010 (N=1801) N (%)	FY2012 (N=1840) N (%)
Sex			
Male	663 (49.6)	937 (52.0)	941 (51.1.)
Female	673 (50.4)	864 (48.0)	899 (48.9)
Race			
White	395 (43.7)	825 (57.9)	872 (59.5)
Non-White	509 (56.3)	599 (42.1)	594 (40.5)
Ethnicity			
Hispanic	292 (32.3)	494 (41.6)	872 (59.5)
Non-Hispanic	612 (67.7)	694 (58.4)	594 (40.5)
Primary Language			
English	1095 (82.0)	1481 (82.2)	1498 (81.4)
Spanish	125 (9.4)	169 (9.4)	212 (11.5)
Other/unknown	115 (8.6)	151 (8.4)	130 (7.1)

Data source 3: Billing and demographic data extracted from Partners Healthcare System data warehouse for claims for WCV and BH screens and demographic data for all 4-17-year-old patients from four MGH-affiliated outpatient pediatric clinics from 7/1/2014-12/31/2016. All patients had Medicaid health insurance, mean age at first WCV was 9.4 years and 52% were male. It may be important to note that in these clinics, patients with Medicaid made up about half of the cases overall and with individual clinics ranging from 25% to 74% Medicaid and virtually all of the rest of the cases having commercial insurance. To keep Data source 3 as comparable as possible to Data sources 1 and 2 (and because there was no mandate to screen or bill for screening for children with commercial insurance), only patients with Medicaid were included in the analyses reported here. Because MGH pediatric clinics switched to a new EMR in 2016, we were not able to do chart reviews on WCV from earlier years.

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.

Data source 1 provides information on all CBHI WCV and BH screens in the state of Massachusetts in order to establish rates of screening overall and by age group and region over a period of close to a decade. Data source 2 was used to examine the percentage of screens using the PSC in particular as well as differences by region of state for three cohorts. Data source 3 was used to provide claims data and chart reviews at the individual case and clinic level and to assess validity and reliability of the PSC as a process measure.

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

To measure patient-level sociodemongraphic variables, we used patient gender, race, ethnicity, and language in Data source 2, and age and gender in Data source 3. These variables were derived from the administrative claims data and/or chart review from all cases in Data source 2 and from administrative data in Data source 3.

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*)

Before presenting evidence for the reliability and validity of the PSC as a process measure, we will briefly summarize the evidence for the PSC instrument.

Reliability and validity of the PSC instrument

The reliability and validity of the PSC at the instrument level have been established in many studies over a period of more than three decades. Two recent systematic reviews have summarized these studies (Lavigne, Meyers, & Feldman, 2016; Pourat, Zima, Marti, & Lee, 2017) and generally confirmed acceptable/high levels of reliablity and validity. A brief summary of specific early and recent studies establishing the validity and reliability of the PSC as an instrument is presented in the next four paragraphs. The reliability/validity of using the CPT code in administrative data or EMR notes to signify the completion of an actual PSC screen are reported in the section that follows.

Early studies established the validity of PSC case/not case coding in outpatient pediatric samples by demonstrating that a PSC score ≥ 28 had a sensitivity of 95% and a specificity of 68% using the clinician-rated Children's Global Assessment Scale (CGAS) score of < 71 as a gold standard in mixed SES (Jellinek et al., 1988) and inner city (Murphy, Arnett, Bishop, Jellinek, & Reede, 1992) samples (sensitivity =.88; specificity =1) and with presence/ absence of a psychiatric diagnosis (sensitivity =.75; specificity=.75) (Murphy, Reede, Jellinek, & Bishop, 1992). Early studies also established the internal reliability of the PSC with Cronbach alpha of .86 (Jellinek, Murphy, & Burns, 1986), and test/retest reliability of two PSCs, one week apart with 91% agreement and a kappa of .69 (Jellinek et al., 1986).

The acceptability and feasibility of the PSC as a routine screen in outpatient pediatric practice were demonstrated in a large (N=20,000+) national probability sample that used the PSC as a proxy for overall psychosocial risk vs not risk (Kelleher et al., 1997). The same sample was used to derive and validate three subscales (attention, internalizing, and externalizing problems) which showed sensitivities of 77-87% and specificities of 68-80% when compared to other widely used measures of problems in these areas (Gardner et al., 1999) and acceptable rates of reliability with Cronbach alphas of .79-.83 on these subscales.

Over the intervening decades, the validity and reliability of the PSC instrument have been demonstrated in many studies. To review some of the more recent, a 6,526 case Florida Medicaid sample (Boothroyd & Armstrong, 2010) reported a sensitivity of .77 and specificity of .82 against presence/absence of mental health disability and a Cronbach alpha of .94 and an even more recent study (Murphy et al., 2016) reviewed normative data for the PSC-17 in a very large (80,000+ case) national outpatient pediatric sample and reported an overall risk rate (12%) and subscale scores that were comparable to rates published in the Kelleher et al, 1997 national sample ~ 20 years earlier, with comparable reliability (Cronbach alpha =.89 and time to time reliability = .85) as well.

The PSC has also been validated as an instrument for assessing change in relationship to treatment (Guzmán et al., 2015; Kamin et al., 2015; McCarthy et al., 2016; Murphy et al., 2015; Murphy et al., 2011; Murphy et al., 2012; Stein et al., 2003).

Reliability and validity of the PSC as a process measure

To test the reliability of the coding of the administration of a PSC tool, we performed chart reviews to ascertain whether the CPT code (96110) used to bill for screening corresponded with evidence in the progress note for the WCV that a PSC or other approved screen had actually been given. Using the presence/absence of evidence of PSC administration in the visit note as the gold standard, we calculated the sensitivity, specificity, and kappa of the CPT code charge for the screen. In order to test the reliability of other critical data elements, we also coded: 1) whether there was a progress note or other evidence of an encounter in the chart documenting that a WCV had occurred on the date it was billed and 2) whether the code used for the WCV was correct based on the age of the child. In addition, we evaluated the inter-rater reliability of all of these assessments by having a second rater code (blind to first reviewer's coding) one third of the cases on all three variables.

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*)

Reliability testing for CPT billing codes compared to chart review as gold standard

From the dataset of all WCV with 4-17-year-old children covered by Medicaid from the four MGH-affiliated outpatient pediatric clinics for calendar year 2016, we were able to access complete EMR data from three clinics resulting in a sample of 6,462 children. For each of these, the dataset included presence or absence of the CPT billing code used for psychosocial screening (96110) on the same day as the well child visit, age and several other demographic and billing variables. In each clinic, 30 WCV were selected and a research assistant opened the electronic medical record for each patient, checking to see whether the progress note for the visit documented that a CBHI approved psychosocial screen had been given, either by the mention of a behavioral health screen and a score, or by finding a PDF of the PSC or another screen in the patient's chart. In each clinic, we coded the first 15 and the last 15 WCV of 2016. Cases marked as restricted access (highly confidential) were skipped and replaced with the next case that matched the selection criteria.

Out of the 90-case sample, 52 (57.8%) of the cases had a 96110 billing code on the date of their well-child visit in 2016 and 38 (42.2%) of the cases did not have a 96110 code. Using documentation of a screen in the chart as the gold standard, there were 59 cases in which we found documentation of a screen in the chart. The 96110 coding correctly identified 52 of these patients (88.1%) as having had a screen and of the 31 patients with no screen in their charts, all 31 (100.0%) were correctly identified as not screened (no 96110 code). Kappa was .84.

The Kappa statistic has the following interpretation:

0.00=Poor; 0.01 – 0.20=Slight; 0.21 – 0.40=Fair; 0.41 – 0.60=Moderate;

0.61 – 0.80= Substantial; 0.81 – 0.99=Almost perfect agreement

Therefore, in this sample, using a billing code to establish the presence of psychosocial screen during a WCV was found to have a very high level of validity.

We also checked for other key data elements using the same method. The first coder also assessed whether there was documentation in the chart that a WCV had occurred on the date billed and found that this coding was perfect: all 90 cases (100%) had documentation that a WCV had been completed on the date that it was billed. Since there are 10 different CPT codes for WCV depending on the age of the child, we also checked 30 cases to compare the age of the child with the specific WCV CPT code used and found that the age appropriate WCV CPT code had been used in 100% of the cases.

To evaluate previous reports (Romano-Clarke et al., 2014; Savageau et al., 2017; Savageau et al., 2016) that about 40% of all screens done in CBHI (and 2/3 to ¾ of the screens for 4-17 year olds) were PSC's, we checked the name of the screen listed in the visit note. In the 90 case chart review sample, the first coder found that of the 59 cases in which there was documentation of a screen, 54 (92%) reported that the PSC was the screen that had been given, 3 (5.1%) mentioned another CBHI approved screen (the Parents Evaluation of Developmental Status (PEDS)) and 2 (3.4%) did not give the name of the screen. Since the PEDS can be used for children from 6 months to 8 years of age, the choice of the screen in these three cases was appropriate and since the PSC has been well-validated for children who are 4-17 years old and since the coder found that all charts that mentioned a specific screen were of children who were within this age range, this coding also showed that the choice of instruments was age appropriate in 100% of cases.

As noted above, we assessed the inter-rater reliability of this medical record coding by having a second research assistant code 30 of the 90 charts (10 from each clinic) selected at random. For the 96110-by-chart-review coding, the ratings of the two RA's agreed on the presence or absence of a screen in 28/30 (93%) of the charts and kappa was .86. In the same 30-case subsample, the second coder also found evidence of a well child visit on all of the dates billed (100% agreement with billing codes and with the first coder). The two coders agreed on the name of the specific screen used in 94% of the cases.

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?*)

This coding showed that that the presence of the PSC or other brief BH screen during WCV could be reliably coded from the presence of the CPT 96110 code in administrative claims data. This approach to coding was also validated by the finding that 100% of the WCV that were billed for were documented by EMR notes from the same day and that the age codes for the WCV were congruent with the age of each child. High levels of interrater reliability for all of these types of coding were also documented. The presence of the PSC as the named screen in 92% of these cases confirmed previous reports (Romano-Clarke et al., 2014; Savageau et al., 2016) that the great majority of the screens used with 4-17 year old's in the CBHI were PSC's.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☐ **Performance measure score**

☐ **Empirical validity testing**

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

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

2b1.3. What were the statistical results from validity testing?

The validity of the PSC instrument has been summarized above in the first part of section **2a2.2.** Since in many respects, evidence for the validity of a process measure is the same as evidence for its reliability, the second part of section 2a2.d can be taken as evidence for the validity of the PSC as a performance measure.

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

Coding the administration of the PSC or other brief BH screen during a WCV from administrative claims data is valid.

2b2. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section* 2b3

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

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with _risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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

N/A

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

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

N/A

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

N/A

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

N/A

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

N/A

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

Statewide summary data from Massachusetts (Data source 1) showed large differences in rates of screening in different regions of the state that were consistent over the entire 9+ years of CBHI. In a chart review subsample of these cases from three years (Data source 2) we were able to access the actual data and test the differences in rates of screening by region using chi square statistics (Savageau et al., 2017).

In a sample for which we had individual case data on WCV by clinic, we were able to compare rates of screening in four MGH-affiliated outpatient pediatric practices clinics using chi square.

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

As also presented in Table 1 in our Main Form, Data source 1 shows clear and consistent differences between the regions of the state in rates of screening for all age groups over the entire 9+ year period. The Western region is always highest, followed by Southeast and Central Massachusetts. Metro West and Boston are always the lowest. Since we did not have access to the actual data, we could not test these differences but they are clearly large and would undoubtedly be statistically significant.

Region	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017*
Western	38.14%	66.13%	75.34%	79.16%	79.79%	80.33%	81.62%	83.76%	81.27%	85.27%
Central	32.61%	51.99%	60.61%	63.22%	67.23%	70.59%	71.09%	70.90%	72.33%	72.31%
NE	32.70%	56.36%	60.93%	60.31%	55.35%	65.08%	67.16%	68.43%	71.46%	67.38%
Metro W	25.89%	47.45%	51.92%	51.80%	57.72%	60.47%	59.83%	54.90%	49.84%	47.72%
SE	36.88%	61.89%	70.57%	75.94%	78.18%	80.19%	80.66%	77.87%	75.91%	73.95%
Boston	21.55%	45.91%	55.28%	57.26%	60.53%	64.49%	67.57%	65.12%	63.27%	58.76%

We did have access to individual case data in Data source 2 and were able to show that the differences between regions were statistically significant, as shown below (Table 4 in our Main Form).

Table 4. Rates of Formal Screening for each year for children ages 4-17

	Formal Screen in 2007		**Formal Screen in 2010**		**Formal Screen in 2012**	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**
Region		*[1]		***[2]		***[3]
Western	139 (95.2%)	7 (4.8%)	22 (11.9%)	163 (88.1%)	33 (15.2%)	184 (84.8%)
Central	64 (94.1%)	4 (5.9%)	30 (23.3%)	99 (76.7%)	31 (23.9%)	99 (76.1%)
Northeast	125 (91.2%)	12 (8.8%)	63 (37.5%)	105 (62.5%)	71 (39.0%)	111 (61.0%)
Metro West	84 (94.4%)	5 (5.6%)	54 (32.5%)	112 (67.5%)	62 (37.4%)	104 (62.7%)
Southeastern	197 (98.5%)	3 (1.5%)	52 (23.0%)	174 (77.0%)	27 (11.1%)	217 (88.9%)
Boston	149 (98.0%)	3 (2.0%)	51 (25.4%)	150 (74.6%)	33 (18.2%)	148 (81.8%)

2007: [1]x^2=13.33 p=.021 **2010**: [2]x^2=36.33 p<.001 **2012:** [3]x^2=75.21 p<.001

In Data source 3, we had a large sample of billing data on BH screening at the individual case level from four clinics. As shown in Table 5 below, there was one clinic that was not screening at all, one that was screening at a relatively low rate (24.1%) and two that were screening at high rates (86.1% and 88.1%), a difference that was statistically significant (p < .001).

It may also be important to reiterate at this point that a study (Hacker, et al, 2016) that compared Medicaid (MAX) claims data from the state of Massachusetts with claims from the state of California over the same 18 month period from 2007 to 2009 found significant differences (~5 fold) in rates of 96110 screening after the start of mandatory screening in MA, suggesting not only a large and continuing performance gap in different states but also evidence that low rates of screening could be overcome on a statewide basis.

Similarly, with reference to the Data source 1 table cited on the previous page, it may also be important to note that by 2017, one of the six regions of Massachusetts had surpassed the 80%-of-WCV-with-screens benchmark sought by CBHI, two other regions were in the low to mid 70% range (which was close to the benchmark) and two other regions far below the bench mark (in the 40-50%% range), again illustrating the large differences in rates of screening in different locales and the large room for improvement in some of them. The same points could be made with reference to Table 5 (below) which shows two clinics in the 80%+ range, one in the 20% range, and one that was not screening at all…underscoring both that high rates of screening can be achieved and that some clinics have not achieved them.

Table 5: Rates of screening in four MGH affiliated outpatient pediatric clinics in 2016

			2016 (n=6,801 children): January 1 2016 – December 31 2016		
Variable	**Clinic A**	**Clinic B**	**Clinic C**	**Clinic D**	**p-value for chi square**
Billed for MH screen	86.1% (3422)	88.1% (1215)	0.0% (0)	24.1 (106)	<.0001

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

Differences like those mentioned in 2b4.2 are both statistically and meaningfully different, suggesting that this measurement strategy has the power to detect differences that are important as well as measurable.

_____

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

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

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

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

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

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

_____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Missing data (in this case, PSC screening scores) is not an issue since those patients who have WCV and are not assessed in the measurement period remain in the denominator.

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

Patients who are seen and not assessed with the PSC during the assessment period are included in the denominator.

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

Missing data is not an issue for this measure as constructed.

References

Boothroyd, R. A., & Armstrong, M. (2010). An examination of the psychometric properties of the Pediatric Symptom Checklist with children enrolled in Medicaid. Journal of Emotional and Behavioral Disorders, 18(2), 113-126.

Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., . . . Chiapetta, L. (1999). The PSC-17: A brief pediatric symptom checklist with psychosocial problem subscales. Ambulatory Child Health, 5, 225-236.

Guzmán, J., Kessler, R. C., Squicciarini, A. M., George, M., Baer, L., Canenguez, K. M., . . . Murphy, J. M. (2015). Evidence for the effectiveness of a national school-based mental health program in Chile. Journal of the American Academy of Child & Adolescent Psychiatry, 54(10), 799-807. e791.

Jellinek, M. S., Murphy, J. M., & Burns, B. J. (1986). Brief psychosocial screening in outpatient pediatric practice. The Journal of Pediatrics, 109(2), 371-378.

Jellinek, M. S., Murphy, J. M., Robinson, J., Feins, A., Lamb, S., & Fenton, T. (1988). Pediatric Symptom Checklist: screening school-age children for psychosocial dysfunction. The Journal of Pediatrics, 112(2), 201-209.

Kamin, H. S., McCarthy, A. E., Abel, M. R., Jellinek, M. S., Baer, L., & Murphy, J. M. (2015). Using a brief parent-report measure to track outcomes for children and teens with internalizing disorders. Child Psychiatry & Human Development, 46(6), 851-862.

Kelleher, K. J., Childs, G. E., Wasserman, R. C., McInerny, T. K., Nutting, P. A., & Gardner, W. P. (1997). Insurance status and recognition of psychosocial problems: a report from the Pediatric Research in Office Settings and the Ambulatory Sentinel Practice Networks. Archives of pediatrics & adolescent medicine, 151(11), 1109-1115.

Lavigne, J. V., Meyers, K. M., & Feldman, M. (2016). Systematic Review: Classification Accuracy of Behavioral Screening Measures for Use in Integrated Primary Care Settings. Journal of Pediatric Psychology, 41(10), 1091-1109. doi: 10.1093/jpepsy/jsw049

McCarthy, A., Asghar, S., Wilens, T., Romo, S., Kamin, H., Jellinek, M., & Murphy, M. (2016). Using a Brief Parent-Report Measure to Track Outcomes for Children and Teens with ADHD. Child Psychiatry & Human Development, 47(3), 407-416.

Murphy, J. M., Arnett, H. L., Bishop, S. J., Jellinek, M. S., & Reede, J. Y. (1992). Screening for psychosocial dysfunction in pediatric practice: A naturalistic study of the Pediatric Symptom Checklist. Clinical pediatrics, 31(11), 660-667.

Murphy, J. M., Bergmann, P., Chiang, C., Sturner, R., Howard, B., Abel, M. R., & Jellinek, M. (2016). The PSC-17: subscale scores, reliability, and factor structure in a new national sample. Pediatrics, e20160038.

Murphy, J. M., Blais, M., Baer, L., McCarthy, A., Kamin, H., Masek, B., & Jellinek, M. (2015). Measuring outcomes in outpatient child psychiatry: Reliable improvement, deterioration, and clinically significant improvement. Clinical child psychology and psychiatry, 20(1), 39-52.

Murphy, J. M., Masek, B., Babcock, R., Jellinek, M., Gold, J., Drubner, S., . . . Hacker, K. (2011). Measuring outcomes in outpatient child psychiatry: The contribution of electronic technologies and parent report. Clinical child psychology and psychiatry, 16(1), 146-160.

Murphy, J. M., Reede, J., Jellinek, M. S., & Bishop, S. J. (1992). Screening for psychosocial dysfunction in inner-city children: further validation of the Pediatric Symptom Checklist. Journal of the American Academy of Child & Adolescent Psychiatry, 31(6), 1105-1111.

Murphy, M., Kamin, H., Masek, B., Vogeli, C., Caggiano, R., Sklar, K., . . . Gold, J. (2012). Using brief clinician and parent measures to track outcomes in outpatient child psychiatry: longer term follow-up and comparative effectiveness. Child and Adolescent Mental Health, 17(4), 222-230.

Pourat, N., Zima, B., Marti, A., & Lee, C. (2017). California Child Mental Health Performance Outcomes System: Recommendation Report. Unpublished Manuscript. UCLA Center for Health Policy Research, 1-298.

Romano-Clarke, G., Tang, M. H., Xerras, D. C., Egan, H. S., Pasinski, R. C., Kamin, H. S., . . . Murphy, J. M. (2014). Have rates of behavioral health assessment and treatment increased for Massachusetts children since the Rosie D. decision? A report from two primary care practices. Clinical pediatrics, 53(3), 243-249.

Savageau, J. A., Keller, D., Simons, J., Lucke, C., Jellinek, M., & Sherwood, E. (2017). Assessing Disparities in Mental Health Screening and Services Before and After Implementation of an Innovative Statewide Program for Children with Medicaid. Presented at Pediatric Academic Societies Meeting, San Francisco, CA, May 7.

Savageau, J. A., Keller, D., Willis, G., Muhr, K., Aweh, G., Simons, J., & Sherwood, E. (2016). Behavioral Health Screening among Massachusetts Children Receiving Medicaid. The Journal of Pediatrics, 178, 261-267.

Stein, B. D., Jaycox, L. H., Kataoka, S. H., Wong, M., Tu, W., Elliott, M. N., & Fink, A. (2003). A mental health intervention for schoolchildren exposed to violence: A randomized controlled trial. Jama, 290(5), 603-611.

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*) Update this field for **maintenance of endorsement**.

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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

The degree of difficulty for data collection varies considerably depending on the system used. In a system like CBHI where all individuals have the same insurer and screening is mandatory and documented with a billing code, data collection carries little burden in terms of time or cost. In a query of billing data, the clinics or other entities and date ranges to be assessed are specified and then the dates for all WCV checked for whether a billing code for screening was submitted for the same date.

Even in a system with more than one insurer, data collection can also be non-burdensome if all entities use a common electronic medical record (EMR) and if specific fields for noting the presence of the screen are agreed to in advance and if use of the billing code for screening is required and complied with internally. For example, in one of the earliest evaluations of the pediatric psychosocial screening as a process measure, Hacker and her associates describe a system in which PSC forms were administered on paper and clinicians noted the exact score in a single field in the EMR (Hacker, Williams, Myagmarjav, Cabral, & Murphy, 2009). A more elaborate version of this system is being used in the outpatient pediatric clinic on at Boston Medical Center. Parents complete the PSC on paper and then medical assistants enter the answers to each item into a template in the EMR (Epic) where scores are computed and data are displayed in flowsheets along with data on height, weight, blood pressure, etc. The outpatient pediatric clinics at Massachusetts General Hospital follow a similar procedure in that PSC scores and items appear in a flowsheet in Epic but are entered by parents who complete the PSC in the waiting room on iPads or at home over a patient portal. There are also standalone electronic systems like CHADIS that can be used independent of EMR that keep track of WCV and which of them included a completed PSC.

For sites that do not mandate the use of a separate billing code for each screen and that lack a common EMR or standalone system, it is still possible to use claims data to create a list of WCV and then to open each electronic or paper chart to look in specific areas for documentation that a PSC was given (Romano-Clarke et al., 2014). Although feasible, this is a much more difficult and time consuming method.

The reference list is included in the attached appendix.

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 fees, licenses or other requirements needed to use any aspect of the measure or the instrument.

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.

Specific Plan for Use	Current Use (for current use provide URL)
	Payment Program Massachusetts Medicaid (MassHealth) http://www.mass.gov/eohhs/consumer/insurance/cbhi/cbhi-screening/ Professional Certification or Recognition Program American Board of Pediatrics Maintenance of Certification https://www.abp.org/content/maintenance-certification-moc Quality Improvement (external benchmarking to organizations) Massachusetts Medicaid (MassHealth) PCC Plan http://www.mass.gov/eohhs/docs/masshealth/provider-services/forms/pcc-handbook.pdf

4a1.1 For each CURRENT use, checked above (update for <u>maintenance of endorsement</u>), provide:

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

Public Reporting:
Program: Behavioral Health (BH) Screening Cumulative Quarterly Report
Sponsor: MassHealth
Purpose: To demonstrate compliance with federal EPSDT regulations requiring screening and the mandate to measure and document progress toward the goal of having 100% of all pediatric well child visits receiving a screen that were a part of the consent decree in the Rosie D vs Romney lawsuit, Massachusetts state Medicaid officials are required to post quarterly reports of the number of behavioral health screenings and well child visits (Behavioral Health Screening Cumulative Quarterly Report BHSCQR) on the MassHealth website.
Level of Measurement and Setting: The geographical area is the state of Massachusetts. Accountable entities include all providers of WCV to MassHealth members and results are reported at the state and regional level.
Professional Certification or Recognition Program:
Program MOC Part 4 Certification
Sponsor: American Board of Pediatrics
Purpose: To maintain of certification with the ABP, pediatricians must complete a number of requirements every 5 years. A system of record keeping that tracks the percentage of well child visits receiving the PSC over a multi-month period as the basis of the MOC quality improvement has been used by the Pediatric Specialty Group of Maine Medical Partners. CHADIS, a commercial software company provides a software system for keeping track of behavioral health screening (96110) to meet this requirement.
Level of Measurement and Setting: The geographical area is variable. Accountable entities are pediatricians/pediatric practices. Level of reporting is usually at the clinician and practice level.
Quality improvement with Benchmarking (external benchmarking to multiple organizations)
Program: PCC Profile Report
Sponsor: Massachusetts Medicaid (MassHealth) Primary Care Clinician (PCC) Plan.
Purpose: The MassHealth PCC plan covers about 170,000 children, about half of those with Medicaid coverage in the state of Massachusetts. The PCC plan monitors and provides biannual feedback to all clinics with a Profile Report that shows clinic rates of behavioral health screening in WCV along with rates of other psychosocial and medical quality measures for that clinic and in comparison with benchmarks from the state as whole.
Level of Measurement and Setting: The geographical area is the state of Massachusetts. Accountable entities include all pediatric practices with 180 or more MassHealth members in enrolled in the PCC Plan (membership ~170,000 children). Level of reporting is at clinic and state.

4a1.2. 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?*)

N/A

4a1.3. 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.*)

N/A

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

According to the PCC Plan Provider Handbook (http://www.mass.gov/eohhs/docs/masshealth/provider-services/forms/pcc-handbook.pdf) "The PCC Profile Report provides information on selected clinical measures, such as pediatric behavioral health, well-child care, and women's cancer screening, that may be used to improve health care delivery and, ultimately, the health outcomes of PCC Plan members. Most of the profile measures display rates of performance for a PCC's practice as well as rates for each service location, if applicable, and for the PCC Plan as a whole. Prior rates for PCC practices and the PCC Plan are also presented to show trended rates for these clinical indicators. Summary data are provided to help with the identification of barriers to care." (page 24). For sites with at least 180 PCC Plan members… a Regional Network Managers (RNMs) … visits in order to review with the PCC the PCC Profile Report, the PCC Reminder Report, and the PCC Care Monitoring Registries. By reviewing the reports and discussing how rates reflect the PCC practice's performance, RNMs assist practices to identify areas for improvement and to develop action plans to improve performance and the delivery of high-quality health care to members". (page 23).

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Not known

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Not known

4a2.2.2. Summarize the feedback obtained from those being measured.

Not known

4a2.2.3. Summarize the feedback obtained from other users

Not known

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Not known

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

Table 3 (section 1b.2) shows the increase in rates of screening in the full sample of 3-17 year olds over the first 9.25 years of CBHI. Table 1 shows that these increases were present in all regions of the state. Rates of screening for all ages increased dramatically, 14-fold, from ~ 5% of all WCV to > 70% over the first three years of the program and have been sustained at that level ever since. For 3-17 year olds who were screened primarily with the PSC, the number of visits screened rose steadily over the first 7 years of the program (from ~80,000 to over 218,000 per year and from 40% to 79%). Almost as important as the large increase in screening is the fact that it has been sustained at over 70% for the past eight years.

Another demonstration of the increase in screening comes from a two state comparison (Hacker et al., 2016). involving almost 10 million well child visits over 4 years. In the year prior to the start of CBHI (2007), both Massachusetts and California were billing for formal BH screens in less than 2 per thousand enrolled youth per month. In the first nine months of 2008 the rates of the BH screening had risen to 13 per thousand children in Massachusetts while remaining at the same level in California.

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

There have been no reports of unintended negative consequences to individuals or populations.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

Benefits include:

-Increasing widespread use of a simple but effective PRO tool that can be used for screening, diagnosis and the monitoring of treatment outcomes for psychosocial problems (California and other states).

-Increased national use of the measure (PSC is being used in the SAMHSA National System of Care Expansion Evaluation and in the state of California child mental health outcomes assessment program), Mental Health America making the PSC and PSC-Y available for free and to tens of thousands of youth or their parents.

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.

Yes

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

0712 : Depression Utilization of the PHQ-9 Tool

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

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

Yes

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.)

The age range for the PHQ-9 (NQF 712) has recently been expanded to include youth 12 to 17 years of age with a diagnosis of depression. The currently submitted measure, the PSC, screens for a broader band of problems (other emotional problems like anxiety as well as other types of problems like attention and behavior) and a larger age range (3-17) than the PHQ-9. Along with the PHQ-9, the PSC is actually one of the specific tools mentioned by the US Preventive Services Task Force as a screen for depression in youth (Forman-Hoffman et al., 2016). Although studies have shown that the PSC identifies about 80% of the youth with depression who are found with the PHQ-9, only about half of the youth with serious psychosocial problems on the PSC are identified with the PHQ-9 (Richardson et al., 2010).The PSC is a representative of a broader class of screening tools (brief broadband psychosocial screens) that are required for use in conjunction with pediatric well child visits in the Massachusetts EPSDT program. Other similar broadband tools are the Strengths and Difficulties Questionnaires and the Child Behavior Checklist. The Massachusetts EPSDT CBHI program provides a short (now 13) list of approved tools (both broad and narrow band) and allows the pediatrician to use the one deemed most appropriate for each case. In a review of nearly 6000 medical charts, Savageau and her associates found that about 40% of all screens were PSC's compared to only about 1% that were PHQ-9's (Savageau et al., 2016; Savageau et al., 2017, May) suggesting that the PSC is at least in the past ten years more widely used by pediatricians in Massachusetts.

The reference list is included in the attached appendix.

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: Appendix_PediatricSymptomChecklist.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Massachusetts General Hospital

Co.2 Point of Contact: J. Michael, Murphy, mmurphy6@mgh.harvard.edu, 617-724-3163-

Co.3 Measure Developer if different from Measure Steward: Massachusetts General Hospital

Co.4 Point of Contact: J. Michael, Murphy, mmurphy6@mgh.harvard.edu, 617-724-3163-

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.

N/A

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision: 10, 2017

Ad.4 What is your frequency for review/update of this measure? continuous/ongoing

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement: ©1988, M.S. Jellinek and J.M. Murphy, Massachusetts General Hospital

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments: N/A

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: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3317

Measure Title: Medication Reconciliation on Admission

Measure Steward: Centers for Medicare & Medicaid Services

Brief Description of Measure: Percentage of patients for whom a designated (PTA) medication list was generated by referencing one or more external sources of PTA medications and for which all PTA medications have a documented reconciliation action by the end of Day 2 of the hospitalization.

Developer Rationale: The Institute for Healthcare Improvement defines medication reconciliation as "the process of creating the most accurate list possible of all medications a patient is taking…and comparing that list against the physician's admission, transfer, and/or discharge orders, with the goal of providing correct medications to the patient at all transition points within the hospital." (Institute for Healthcare Improvement, 2017). While medication reconciliation should occur at all transition points during the inpatient stay, this measure focuses on medication reconciliation on admission because information collected at this transition point is critical to inform treatment decisions during the inpatient stay and at discharge. By collecting adequate information about a patient's PTA medications, recording the information in a single location in the medical record for easy reference, and reconciling this information in a timely manner, clinicians can avoid potentially harmful medication discrepancies. A thorough reconciliation process is important in the IPF setting because pharmacotherapy is a primary form of treatment for patients with severe psychiatric illnesses and the accuracy of self-reported PTA medications may be compromised by severe psychiatric symptoms.

Studies in both the psychiatric and non-psychiatric settings have found that medication discrepancies are present in more than half of medical records for inpatient stays. (Brownlie, 2014; Cornish, 2005). There is evidence to suggest that most medication discrepancies in inpatient medical records result from the failure to collect and reconcile PTA medications. The Multicenter Medication Reconciliation Quality Improvement Study (MARQUIS), which was conducted in six U.S. hospitals, reported an average of 3.35 unintentional medication discrepancies per patient with most medication discrepancies (2.12 per patient) resulting from failure to accurately identify the patient's PTA medications (Salanitro, 2013). The Medications At Transitions and Clinical Handoff (MATCH) study evaluated 651 inpatient stays and found that as many as 85% of admissions with medication errors had errors that originated from incomplete collection of the medication history (Gleason, 2010).

To reduce discrepancies that result from inadequate collection and reconciliation of PTA medications, the Medication Reconciliation on Admission measure is constructed to align with the two elements of performance of The Joint Commission's National Patient Safety Goal (NPSG.03.06.01) on medication safety that are relevant to the admission process (The Joint Commission, 2016). These elements are:

- Obtain information on the medications the patient is currently taking when he or she is admitted to the hospital or is seen in an outpatient setting. This information is documented in a list or other format that is useful to those who manage medications.
- Compare the medication information the patient brought to the hospital with the medications ordered for the patient by the hospital in order to identify and resolve discrepancies.

Citations

* Brownlie, K., Schneider, C., Culliford, R., Fox, C., Boukouvalas, A., Willan, C., & Maidment, I. D. (2014). Medication reconciliation by a pharmacy technician in a mental health assessment unit. Int J Clin Pharm, 36(2), 303-309. doi:10.1007/s11096-013-9875-8

*Cornish, P. L., Knowles, S. R., Marchesano, R., Tam, V., Shadowitz, S., Juurlink, D. N., & Etchells, E. E. (2005). Unintended medication discrepancies at the time of hospital admission. Arch Intern Med, 165(4), 424-429. doi:10.1001/archinte.165.4.424

*Gleason, K. M., McDaniel, M. R., Feinglass, J., Baker, D. W., Lindquist, L., Liss, D., & Noskin, G. A. (2010). Results of the Medications at Transitions and Clinical Handoffs (MATCH) study: an analysis of medication reconciliation errors and risk factors at hospital admission. J Gen Intern Med, 25(5), 441-447. doi:10.1007/s11606-010-1256-6

*Institute for Healthcare Improvement. (2017). Medication reconciliation to prevent adverse drug events. Retrieved from http://www.ihi.org/Topics/ADEsMedicationReconciliation/Pages/default.aspx

*Salanitro, A. H., Kripalani, S., Resnic, J., Mueller, S. K., Wetterneck, T. B., Haynes, K. T., . . . Schnipper, J. L. (2013). Rationale and design of the Multi-center Medication Reconciliation Quality Improvement Study (MARQUIS). BMC Health Serv Res, 13, 230. doi:10.1186/1472-6963-13-230

*The Joint Commission. (2016). National patient safety goals effective January 1, 2017: Hospital Accreditation Program. Retrieved from https://www.jointcommission.org/assets/1/6/NPSG_Chapter_HAP_Jan2017.pdf

Numerator Statement: Number of patients for whom a designated Prior to Admission (PTA) medication list was generated by referencing one or more external sources of medications and for which all PTA medications have a documented reconciliation action by the end of Day 2 of the hospitalization when the admission date is Day 0.

Denominator Statement: All patients admitted to an inpatient facility from home or a non-acute setting.

Denominator Exclusions: The measure applies two exclusion criteria to ensure that it is feasible to complete the medication reconciliation process on admission to the IPF:

1. Patients transferred from an acute care setting

2. Patient admissions with a length of stay less than or equal to 2 days

Measure Type: Process

Data Source: Paper Medical Records

Level of Analysis: Facility

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

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a structure, process or intermediate outcome 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. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

- Systematic Review of the evidence specific to this measure? ☒ Yes ☐ No
- Quality, Quantity and Consistency of evidence provided? ☒ Yes ☐ No
- Evidence graded? ☐ Yes ☒ No

Evidence Summary

- The developer provides a logic model.
- The developer includes two systematic reviews for evidence:
    - Hospital-Based medication reconciliation practices: A systematic review. *Archives of Internal Medicine*. 2012. Of the 26 studies identified in the review – 6 were rates as good quality, 5 as fair, and 15 as poor using USPSTF criteria. **No overall grade was provided.**
    - The Joint Commission. (2016). National patient safety goals effective January 1, 2017: Hospital Accreditation Program. **No grade was provided.**

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence relevant enough considering measure focus on reconciliation at admission only?*

Guidance from the Evidence Algorithm

Process measure based on systematic review (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: moderate; Consistency: high (Box 5) → Moderate (Box 5b) → Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

RATIONALE:

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

Maintenance measures – increased emphasis on gap and variation

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

The developer includes rationale for this measure based on the importance of reconciliation process in the Inpatient Psychiatric Facility (IPF) setting because pharmacotherapy is a primary form of treatment for patients with severe psychiatric illness and the accuracy of self-reported PTA medications may be compromised by severe psychiatric symptoms.

The developer provides performance data from nine IPFs who participated in field testing the measure.

- Each of the nine IPFs were asked to abstract information from 100 patient admissions that met testing criteria using one of the two sampling approaches: (1) selection of the most recent admissions; or (2) random selection of admissions.
- Average measure score was 50% with a standard deviation of 32% and ranged from 7% - 98% across the nine facilities.

IPF ID	Measure Score (%)	95% Confidence Interval
IPF 1	68	59,77
IPF 2	18	10, 26
IPF 3	77	69,85
IPF 4	88	82, 94
IPF 5	30	21, 39
IPF 6	7	2, 12
IPF 7	43	33, 53
IPF 8	98	95, 100
IPF 9	18	10, 26

Disparities

Developer provides age, gender, race and ethnicity analyses in testing results.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence

Comments:

** The evidence basis is only moderate--a number of poor or mediocre trials; significant extrapolation from several positive trials; and varying definitions of medication reconciliation under differing circumstances.

** The evidence reviewed makes it hard to discern the importance of the observed discrepancies to patients. Many of the studies are of poor quality and only 5 show outcomes (ADEs) and the specific ADEs are not described (thus their impact is hard to determine from the data presented). This is particularly concerning since both the record-keeping inherent in this measure and the chart review required to calculate perforance could be burdensome for facilities and so the relationship of measurement burden to potentially better patient outcomes is a bit hard to determine.

** This is a patient safety measure and critical to clinical decision-making about treatment during inpatient stays and after discharge. That medication discrepancies are found in half or more of medical records at admission suggests a dangerous trend. Systematic reviews have been conducted but no grades were given to evidence. No other measure focuses on reconciliation at admission. Use of the measure should encourage patient and treatment plan monitoring during treatment and create better understanding of the interaction between pharmacologic and behavioral interventions in a treatment plan. Furthermore, lack of awareness of the PTA medications a patient has been using has

serious morbidity and mortality implications. Outcomes should been viewed as "in-treatment" outcomes---differences in treatment plans etc.
** good evidence to support the measure.
** Good evidence for measure focus given the high level of errors and those errors having some correlation with lack of adequate medication reconciliation on admission to inpatient facilities.
** The developers provide good evidence that mediation reconciliation reduces adverse drug events. No concerns.
** Yes there is evidence of the need of this measure.
** Only 2 studies provided with ungraded evidence. Med Rec has been recognized in prior studies as a method to reduce errors in transition of care processes; however in June 2017 AHRQ Patient Safety Network primer on med Rec states evidence supporting patient benefits from med rec is relatively scanty. Most med rec interactions focus on hospital admission or discharge processes, but most effective and generalizable strategy is unclear. This metric seems to be defining this process using stakeholders to develop workflow - unclear of business process management discussions and integration with analytics to determine if this is the optimal workflow. Concern of laborious process being recommended for measure for admission med rec alone will translate into meaningful improvement in patient outcomes. Many drug-drug interactions identified as potentially serious do not translate to that serious level for all patients. So, again questioning translation of work leading to meaningful improvement in silo. Combining with med rec at discharge to assure the PTA med rec process is used for determining optimal discharge list may provide greater value.

1b. Performance Gap

Comments:
**Yes, there appears to be a gap.
**There are substantial differences between facilities.
**The variability in performance across 10 inpatient psychiatric facilities supports the need for this measure.
**gap demonstrated.
**Great deal of variation among the IPFs.
**The developers collected data from 9 IPFs in 8 states. The results show a large rang in measure performance. No concerns.
** Current performance data provided for a limited number of facility
** Yes, 9 facilities sampled with 100 pts each and were found to have an avg score of 50% with standard deviation of 32%, range: 7-98% across 9 facilities.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing
2b. Validity: Testing; Exclusions**;** Risk-Adjustment**;** Meaningful Differences**;** Comparability**;** Missing Data

Reliability

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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. For maintenance measures – less emphasis if no new testing data provided.

Validity

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. For maintenance measures – less emphasis if no new testing data provided.

2b2-2b6. Potential threats to validity should be assessed/addressed.

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

Complex measure evaluated by Scientific Methods Panel? ☐ **Yes** ☒ **No**

Evaluators: NQF Staff

Evaluation of Reliability and Validity (and composite construction, if applicable):

Link A

Additional Information regarding Scientific Acceptability Evaluation (*if needed*):

N/A

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?*
- *The NQF Staff is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?*

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *The NQF Staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?*

Preliminary rating for reliability:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**
Preliminary rating for validity:	☐ **High**	☒ **Moderate**	☐ **Low**	☐ **Insufficient**

Scientific Acceptability

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.**

Measure Number: 3317

Measure Title: Medication Reconciliation on Admission

RELIABILITY

1. Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?
   NOTE: NQF staff will conduct a separate, more technical, check of eMeasure (eCQM) specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.

   TIPS: Consider the following: 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?

   ☒Yes (go to Question #2)

   ☐No (please explain below, and go to Question #2) NOTE that even though ***non-precise specifications should result in an overall LOW rating for reliability***, we still want you to look at the testing results.
2. Was empirical reliability testing (at the data element or measure score level) conducted using statistical tests with the measure as specified?

TIPS: Check the 2nd "NO" box below if: only descriptive statistics provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level of analysis, patients)

☒Yes (go to Question #4)

☐No, there is reliability testing information, but *not* using statistical tests and/or not for the measure as specified OR there is no reliability testing (please explain below then go to Question #3)

3. Was **empirical VALIDITY testing** of patient-level data conducted?

   ☐Yes (use your rating from data element validity testing – Question #16- under Validity Section)

   ☐No (please explain below and rate Question #11: OVERALL RELIABILITY as INSUFFICIENT and proceed to the VALIDITY SECTION)

4. Was reliability testing conducted with computed performance measure scores for each measured entity?

   TIPS: Answer no if: only one overall score for all patients in sample used for testing patient-level data

   ☒Yes (go to Question #5)

   ☐No (go to Question #8)

5. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? *NOTE: If multiple methods used, at least one must be appropriate.*

   TIPS: Examples of appropriate methods include signal-to-noise analysis (e.g. Adams/RAND tutorial); random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒Yes (go to Question #6) Signal-to-noise analysis

   ☐No (please explain below then go to Question #8)

6. **RATING (score level)** - What is the level of certainty or confidence that the performance measure scores are reliable?

   TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient reliability so that differences in performance can be identified?

   ☒High (go to Question #8)

   ☐Moderate (go to Question #8)

   ☐Low (please explain below then go to Question #7)

7. Was other reliability testing reported?

   ☒Yes (go to Question #8)

   ☐No (rate Question #11: OVERALL RELIABILITY as LOW and proceed to the VALIDITY SECTION)

8. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?

   TIPS: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" see Validity Section Question #15)

   ☒Yes (go to Question #9)

   ☐No (if there is score-level testing, rate Question #11: OVERALL RELIABILITY based on score-level rating from Question #6; otherwise, rate Question #11: OVERALL RELIABILITY as INSUFFICIENT. Then proceed to the VALIDITY SECTION)

9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

TIPS: For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

Answer no if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

☒Yes (go to Question #10) Inter-abstractor agreement (ICC, Kappa)

☐No (if no, please explain below and rate Question #10 as INSUFFICIENT)

10. **RATING (data element)** – Based on the reliability statistic and scope of testing (number and representativeness of patients and entities), what is the level of certainty or confidence that the data used in the measure are reliable?

TIPS: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?

☐Moderate (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as MODERATE)

☒Low (if score-level testing was NOT conducted, rate Question #11: OVERALL RELIABILITY as LOW) Developer states the data element "External Source" has low agreement and a slight - fair Cohen's Kappa score due to medical record documentation practices.

☐Insufficient (go to Question #11)

11. OVERALL RELIABILITY RATING

OVERALL RATING OF RELIABILITY taking into account precision of specifications and all testing results:

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has not been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete]

☐Insufficient (please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required]

Performance score reliability is highly reliable with a sample of 100 records.

The pooled Cohen's Kappa score for the data elements across nine facilities was 0.61 (95% CI: 0.53, 0.69) indicating substantial agreement.

VALIDITY

ASSESSMENT OF THREATS TO VALIDITY

1. Were all potential threats to validity that are relevant to the measure empirically assessed?

TIPS: Threats to validity include: exclusions; need for risk adjustment; Able to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.

☒Yes (go to Question #2)

☐No (please explain below and go to Question #2) [NOTE that even if ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***, we still want you to look at the testing results]

The measure applies two exclusion criteria to ensure that it is feasible to complete the medication reconciliation process on admission to the IPF:

1. Patients transferred from an acute care setting
2. Patient admissions with lengths of stay less than or equal to 2 days

2. Analysis of potential threats to validity: Any concerns with measure exclusions?

TIPS: Consider the following: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? If patient preference (e.g., informed decisionmaking) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the information about patient preference and the effect on the measure is transparent?

☐Yes (please explain below then go to Question #3)

☒No (go to Question #3)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #3)

3. Analysis of potential threats to validity: Risk-adjustment (applies to all outcome, cost, and resource use measures; may also apply to other types of measure)

☒Not applicable (e.g., structure or process measure that is not risk-adjusted; go to Question #4)

a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

b. Are social risk factors included in risk model? ☐Yes ☐No

c. Any concerns regarding the risk-adjustment approach?

*TIPS: Consider the following: If a justification for **not risk adjusting** is provided, is there any evidence that contradicts the developer's rationale and analysis? If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? **If risk adjusted**: Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented? Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)? If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision? Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)? Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

☐Yes (please explain below then go to Question #4)

☐No (go to Question #4)

This is a process measure.

4. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?

☐Yes (please explain below then go to Question #5)

☒No (go to Question #5)

5. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?

☐Yes (please explain below then go to Question #6)

☒No (go to Question #6)

☐Not applicable (go to Question #6)

6. Analysis of potential threats to validity: Any concerns regarding missing data?

☒Yes (please explain below then go to Question #7)

☐No (go to Question #7)

Note that the measure score is largely based on the presence of specific data elements in the medical record. Medication reconciliation within 2 days of admission to an IPF is expected for each patient admission in the measure population. Thus, missing data are considered part of the quality signal in the measure score.

ASSESSMENT OF MEASURE TESTING

7. Was empirical validity testing conducted using the measure as specified and appropriate statistical test?

   Answer no if: face validity; only refer to clinical evidence; only descriptive statistics; only describe process for data management/cleaning/computer programming; testing does not match measure specifications (i.e. data, eMeasure, level, setting, patients).

   ☐Yes (go to Question #10) [NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary. Go to Question #8 **only if** there is insufficient information provided to evaluate data element and score-level testing.]

   ☒No (please explain below then go to Question #8)

8. Was face validity systematically assessed by recognized experts to determine agreement on whether the computed performance measure score from the measure as specified can be used to distinguish good and poor quality?

   TIPS: Answer no if: focused on data element accuracy/availability/feasibility/other topics; the degree of consensus and any areas of disagreement not provided/discussed.

   ☒Yes (go to Question #9)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT)

9. **RATING (face validity)** - Do the face validity testing results indicate substantial agreement that the performance measure score from the measure as specified can be used to distinguish quality AND potential threats to validity are not a problem, OR are adequately addressed so results are not biased?

   ☒Yes (if a NEW measure, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, rate Question #17: OVERALL VALIDITY as INSUFFICIENT; otherwise, rate Question #17: OVERALL VALIDITY as MODERATE)

   ☐No (please explain below and rate Question #17: OVERALL VALIDITY AS LOW)

10. Was validity testing conducted with computed performance measure scores for each measured entity?

    TIPS: Answer no if: one overall score for all patients in sample used for testing patient-level data.

    ☐Yes (go to Question #11)

    ☐No (please explain below and go to Question #13)

11. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?

    TIPS: For example: correlation of the performance measure score on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

    ☐Yes (go to Question #12)

    ☐No (please explain below, rate Question #12 as INSUFFICIENT and then go to Question #14)

12. **RATING (measure score)** - Based on the measure score results (significance, strength) and scope of testing (number of measured entities and representativeness) and analysis of potential threats, what is the level of certainty or confidence that the performance measure scores are a valid indicator of quality?

    ☐High (go to Question #14)

    ☐Moderate (go to Question #14)

    ☐Low (please explain below then go to Question #13)

☐Insufficient

13. Was other validity testing reported?

    ☐Yes (go to Question #14)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as LOW)

14. Was validity testing conducted with patient-level data elements?

    TIPS: Prior validity studies of the same data elements may be submitted

    ☐Yes (go to Question #15)

    ☐No (please explain below and rate Question #17: OVERALL VALIDITY as INSUFFICIENT if no score-level testing was conducted, otherwise, rate Question #17: OVERALL VALIDITY based on score-level rating from Question #12)

15. Was the method described and appropriate for assessing the accuracy of ALL critical data elements? *NOTE that data element validation from the literature is acceptable.*

    TIPS: For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

    Answer No if: only assessed percent agreement; did not assess separately for all data elements (at least numerator, denominator, exclusions)

    ☐Yes (go to Question #16)

    ☐No (please explain below and rate Question #16 as INSUFFICIENT)

16. **RATING (data element)** - Based on the data element testing results (significance, strength) and scope of testing (number and representativeness of patients and entities) and analysis of potential threats, what is the level of certainty or confidence that the data used in the measure are valid?

    ☐Moderate (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as MODERATE)

    ☐Low (please explain below) (if score-level testing was NOT conducted, rate Question #17: OVERALL VALIDITY as LOW)

    ☐Insufficient (go to Question #17)

17. OVERALL VALIDITY RATING

OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.

☐High (NOTE: Can be HIGH only if score-level testing has been conducted)

☒Moderate (NOTE: Moderate is the highest eligible rating if score-level testing has NOT been conducted)

☐Low (please explain below) [NOTE: Should rate LOW if you believe that there are threats to validity and/or threats to validity were not assessed]

☐Insufficient (if insufficient, please explain below) [NOTE: For most measure types, testing at both the score level and the data element level is not required] [NOTE: If rating is INSUFFICIENT for all empirical testing, then go back to Question #8 and evaluate any face validity that was conducted, then reconsider this overall rating.]

Committee Pre-evaluation Comments: Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability – Specifications

Comments:

**There are significant concerns about some elements definition. As noted the external sources was particularly troublesome.
**Reliability generally seems reasonable. Some of the "external sources" for information about medications (e.g., speaking with the patient or a caregiver) seem almost automatic. The measure seems subject to becoming a too simple checkbox measure.
**Specifications are acceptable.
**no concerns
**Not clear that the external source measure should be used or how it contributes to the value of the measure, particularly since an external source can be the patient themselves.
** Is reliable if relying on human interaction to extract. would like to see this done electronically. A check box in the EHR checked by the provider indicating they did the necessary reconciliation would help.
** Test scores indicate overall 87.9% agreement with Cohen's kappa score of 0.61 (equates to substantial agreement; however cutpoint is 0.61 to achieve this, so just made it). External source scored very low, 0.18 (slight agreement). Not sure if this component of measure should be included with so much unreliability.

2a2. Reliability – Testing

Comments:

** Yes, I do. I am not convinced the actual data collected will be the same. It relies on due diligence beyond this measure, and frankly, I am doubtful of this validity which will in turn increase variability and lessen reliability.
** No except as noted above.
** No concerns about reliability and does not need to ben discussed or voted on in Committee.
** No concerns
** My only concern is that the reliability was low on the second criterion that the facilities consult at least one source external to the facilities records to increase comprehensive capture of all active medications ion the PTA medication list. (Cohen's Kappa - 0.18)

2b1. Validity –Testing

2b4-7. Threats to Validity

2b4. Meaningful Differences

Comments:

** Especially the issue of missing data/data fudged.
** no concerns except the reliability issues noted above
** Validity is adequately demonstrated. No need to discuss or vote on in Committee.
** No concerns.
** Validity testing was done based on face validity. No concerns.
** Facilities on Paper records will have a much greater burden of data capture I think. Might have an influence on obtaining the data
** Face validity with 19 of 21 voting members in agreement. Again, I would like to understand the business process management pathway - taking a coordinated view of the performance of all of the process components across the functional organization by which value is exchanged and performance is optimized. This will help me feel more reassured that a new labor-tedious process does not result in unintended negative effect on upstream or downstream processes.

2b2-3. Other Threats to Validity

2b2. Exclusions

2b3. Risk Adjustment:

Comments:

** Having worked with EHRs of many forms, it is often a measure of GIGO--with no meaningful review or questioning of the data recorded. I have experienced this multiple times in my own care and I doubt my experience is unique.
** exclusions seem reasonable.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

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.

- Data Elements are generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
- Some data elements are in defined fields in electronic sources
- The measure was specified to use manually chart-abstracted data from medical records for the following reasons:
    - The setting in which this measure was tested (inpatient psychiatric facilities) primarily used paper records at the time of development.
    - Among IPFs that participate in the IPFQR Program, only about 36% attested to using an EHR system for fiscal year 2016 (CMS, 2016).
- This approach was also selected because many of the data elements are not currently collected in structured, computer-readable fields. The developer anticipates that if this measure were to be implemented, some of the data elements could be collected in structured fields.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 3: Feasibility

3. Feasibility

Comments:

** For most EHRS this is readily available, but not necessarily correct.

** As noted above; because some of the elements are not routinely collected or coded both the record keeping and the chart abstractions seem somewhat burdensome.

** This measure is specified to use manually chart-abstracted data from medical records because so many inpatient psychiatric facilities do not use EHR systems. While feasible for healthcare personnel to collect in the routine course of providing care, the measure poses some additional burden. If put to increased use, the measure may very well push EHR providers to create structured fields for these data elements as has happened with other types of behavioral health data originally specified for manual collection.

** abstraction from medical records.

** Moderate. Routinely generated data but continued reliance on manual paper chart data abstraction makes it more burdensome. relaince stated due to less than 50% psychiatric facilities attest to EHR as of 2016. It's 2018 so perhaps worth reviewing for progress in use of EHR.

** some of the data element are in defined fields in electronic sources but many are not. It is anticipated that once required, some of the elements could be collected electronically.

** Paper chart extraction is cumbersome, but as more facilities use eMRs, should become easier. All fields listed for PTA med rec metric may not be built into eMR, cost to modify eMR. There is a fee paid to data extractors for quality check, so the more burdensome the process the higher the fee.

Criterion 4: Usability and Use

4a. Use (4a1. Accountability and Transparency; 4a2. Feedback on measure)

4a. 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.

4a.1. 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.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

4a.2. Feedback on the measure by those being measured or others. Three criteria demonstrate feedback: 1) those being measured have been given performance results or data, as well as assistance with interpreting the measure results and data; 2) those being measured and other users have been given an opportunity to provide feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- In 2016, the Measure Application Partnership (MAP) recommended that this measure be refined and resubmitted prior to rulemaking because it is currently undergoing field testing.
- In 2017, the composite version of this measure was not recommended for endorsement , by the Behavioral Health Standing Committee. The Committee expressed concern that the evidence was weak for this measure focus, noting that in the 2012 systematic review, only 6 of the 26 studies were rated as good quality. Committee members noted that studies of the medication reconciliation process are usually conducted in acute care facilities, and not in inpatient psychiatric facilities.
- The measure has been revised based on feedback from stakeholders to simplify the measure logic and align data element definitions with similar data elements used by other measures.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *How has the measure been vetted in real-world settings by those being measured or others?*

Preliminary rating for Use: ☒ **Pass** ☐ **No Pass**

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

4b. Usability 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.

4b.1 Improvement. Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.

- This measure is, currently, not publicly reported or in use because this is a new measure.
- The developer anticipates that the implementation of this measure will lead to more standardization in the documentation of medication reconciliation at facilities, which will improve communication across providers and may lead to fewer adverse drug events and better patient outcomes.

Improvement results

N/A this measure has not been implemented.

4b2. Benefits vs. harms. 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).

Unexpected findings (positive or negative) during implementation [unexpected findings]

Potential harms

None reported by the developer

Additional Feedback:

None reported by the developer

***Questions for the Committee**:*

- *How 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?*

Preliminary rating for Usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee Pre-evaluation Comments: Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

** No data as I recall.

** Feedback as given on this measure in 2016 and 2017 and measure has been revised accordingly to simplify the measure logic and align data element definitions with similar data elements used by other measures. Reducing discrepancies that result from inadequate collection and inaccurate recording of PTA medications can have such serious consequences that this measure seems essential.

** The way this is proposed requires manual extraction. Certainly be much better if provider attestation after reviewing could be one entry

** Not in use in present. Presently there is no standardized process for PTA or even discharge med rec. There is a button to click to say med rec was completed, but no standardized process to demonstrate consistent value of med rec.

4b1. Usability – Improvement

Comments:

** Can be useful and probably not much harm, so on the balance, a useful if somewhat flawed measure.

** As noted above, although the measure concept is undeniably important, I have some concerns about quantifying the potential benefit of this measure (the extent to which it improves patient outcomes as opposed just to record keeping) and whether the benefits of the measure outweigh the burden.

** Far from producing harms, use of this measure can provide considerable benefits to treatment planning and implementation during inpatient stays. If patients are monitored during treatment, the benefits of use of this measure should be visible. Related measures do not capture the information this measure captures with the specificity needed. Where appropriate measure has been harmonized with other related measures....no concerns about usability.

** no question that med reconciliation is important and valuable. but i continue to hope for universal medical records which would include medication and allergy profile for optimal patient care experience.

** Other than burden of paper chart abstraction, see no harm and benefit of careful reconciliation is clear.

** Main harm is additional documentation burden but benefits outweigh this concern.

** Should be an eMeasure

** Useful to develop a standardized PTA med rec process. Concerned on whether patient outcome will be improved to the level the high labor process will cost (cost/benefit analysis when trying to implement multiple higher quality initiatives in the hospital). Would be good to connect this with the discharge med rec metric so that the work done to obtain an accurate PTA med list is reviewed again at discharge to assure all PTA drugs are clearly listed as continue, d/c, change along with new drugs due to inpt stay. Not always clear in current discharge drug list output.

Criterion 5: Related and Competing Measures

Related or competing measures

0097 : Medication Reconciliation Post-Discharge

0293 : Medication Information

0553 : Care for Older Adults (COA) – Medication Review

0646 : Reconciled Medication List Received by Discharged Patients (Discharges from an Inpatient Facility to Home/Self Care or Any Other Site of Care)

2988 : Medication Reconciliation for Patients Receiving Care at Dialysis Facilities

Harmonization

- The Measure Developer evaluated existing measures in the NQF portfolio to determine whether the Medication Reconciliation on Admission measure would compete with existing measures and made the following conclusions:
    - NQF measures #0097, #0553, and #2988 are specified for the outpatient setting and the two NQF measures #0293 and #0646 are specified for the inpatient setting focus on communication of information at discharge. The Medication Reconciliation on Admission measure is the only measure that evaluates medication reconciliation on admission to an inpatient facility.
    - The measure has been harmonized with other related/competing measures timeframe specifications and data elements.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: January 10, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

1. Evidence and Performance Gap – 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 sub criteria to pass this criterion and be evaluated against the remaining criteria.***

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

171004_Med_Rec_nqf_evidence_attachment_7.1.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission? Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1a Evidence (subcriterion 1a)

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

Measure Title: Medication Reconciliation on Admission

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

Date of Submission: Click here to enter a date11/1/2017

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete **EITHER 1a.2, 1a.3 or 1a.4** as applicable for the type of measure and evidence.*
- *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.*
- 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.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Standing 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:

- Outcome: [3] Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- 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.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

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 Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

☐ Outcome:

☐Patient-reported outcome (PRO):

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*A PRO-based performance measure is not a survey instrument. Data may be collected using a survey instrument to construct a PRO measure.)*

☐ Intermediate clinical outcome (*e.g., lab value*):

☒ Process: Medication Reconciliation on Admission

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

1a.2 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

The logic model establishing the process-outcome link for this measure concept is listed below. The process steps corresponding to the measure concept are shown in bold. Literature supporting this logic model is provided in Section 4b.

Patient is admitted for inpatient care → **Care team documents on a designated PTA Medication List all medications taken by the patient prior to admission → The PTA medication list is generated using at least one source external to the facility's records to identify the medications taken by the patient prior to admission → Licensed prescriber documents a reconciliation action to continue, discontinue, or modify each medication listed**

on the PTA Medication List by the end of Day 2 of the hospitalization, or if there are no medications on the PTA medication list, the prescriber signs the document by the end of Day 2 of the hospitalization to indicate his/her review of the PTA medication list → Medication errors during the inpatient stay and at discharge are reduced → Adverse drug events are prevented

1a.3 Value and Meaningfulness: IF this measure is derived from patient report, provide evidence that the target population values the measured ***outcome, process, or structure*** and finds it meaningful. (Describe how and from whom their input was obtained.)

This measure is not derived from patient report. However, patients may be a source of PTA medications.

****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4) ****

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

The evidence for this measure includes:

- A systematic review published in 2012 and additional studies identified since the review.
- Standards for Medication Reconciliation put forth in the National Patient Safety Goals by The Joint Commission.

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	**Mueller, S. K., Sponsler, K. C., Kripalani, S., & Schnipper, J. L. (2012). Hospital-based medication reconciliation practices: A systematic review. *Archives of Internal Medicine, 172*(14), 1057-1069. doi: 10.1001/archinternmed.2012.2246**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	A systematic review published in 2012 identified 26 controlled studies related to hospital-based medication reconciliation practices (Mueller, Sponsler, Kripalani, & Schnipper, 2012). This review used the 2007 Institute for Healthcare Improvement definition of medication reconciliation, which is the "process of identifying the most accurate list of all medications a patient is taking…and using this list to provide correct medications for patients anywhere within the health care system." The review concludes that the identified studies "consistently demonstrated a reduction in medication discrepancies (17/17 studies), potential adverse drug events (5/6 studies), and adverse drug events (2/2 studies). Key aspects of successful interventions included intensive pharmacy staff involvement and targeting the intervention to a 'high-risk' patient population." Of note, the systematic review did not discriminate between medication reconciliation at admission, transfer between hospital units, or discharge. *Institute for Healthcare Improvement. (2007) Medication reconciliation review. Retrieved on May 3, 2012 from http: //www.ihi.org/knowledge/Pages/Tools/ MedicationReconciliation Review.aspx
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The systematic review did not provide an overall grade for the body of evidence. Of the 26 studies identified, 6 were rated as good quality, 5 as fair, and 15 as poor, using the United States Preventive Services Task Force (USPSTF) criteria.
Provide all other grades and definitions from the evidence grading system	The USPSTF grades the quality of the overall evidence for a service on a 3-point scale (good, fair, poor): • Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes. • Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes. • Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.
Grade assigned to the **recommendation** with definition of the grade	The authors of the systematic review did not make specific recommendations.
Provide all other grades and definitions from the recommendation grading system	The authors of the systematic review did not provide other grades and definitions.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Twenty-six controlled studies were included in the systematic review. Ten of the studies were randomized controlled trials, three were nonrandomized trials with a concurrent control group, and 13 were pre-post studies. Based on the USPSTF grades, 11 of the 26 studies were graded as good to fair quality.
Estimates of benefit and consistency across studies	The studies in the review "consistently demonstrated a reduction in medication discrepancies (17/17 studies), potential adverse drug events (5/6 studies), and adverse drug events (2/2 studies)".
What harms were identified?	This review did not identify any harms from this intervention.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	To identify additional studies since the systematic review by Mueller et al., we conducted an additional literature review of recent studies that evaluated the impact of (optimized) medication reconciliation-at-admission on the reduction of medication discrepancies, potential adverse drug events (pADEs), or adverse drug events (ADEs). Using PubMed, we searched a Medical Subject Headings (MeSH) of *medication reconciliation* as a search term to retrieve additional studies published from March 2012 to October 2017. A total of 591 studies were identified and 16 studies were considered as relevant via initial independent assessment of titles and subsequent examination of abstracts with or without full-text review. Review of the new 16 studies confirmed the conclusion from the systematic review; performing medication reconciliation at admission significantly decreased medication discrepancies (10/13 studies) (Byrne, Grimes, Jago-Byrne, Galvin, 2017; Hron, et al, 2015; Sherr, et al, 2011; Gimenez-Manzorro, et al, 2015; Curatolo, Gutermann, Devaquet, Roy, Rieutord, et al,2015; Leguelinel-Blache, et al, 2014; Andreoli, et al, 2014; Grimes, et al, 2014; Lea, Barstad, Mathiesen, Mowe, Molden, 2016; Becerra-Camargo, Martinez-Martinez, Garcia-Jimenez, 2013; Cater, et al, 2013; Wang, Biederman, 2012; van den Bemt, van der Schrieck-de Loos, van der Linden, Theeuwes, Pol AG, 2013), potential adverse drug events (2/3 studies)) Grimes, et al, 2014; Wang, Biederman, 2012; Becerra-Camargo, Martinez-Martinez, Garcia-Jimenez, 2014) and adverse drug events (3/3 studies) (Hron, et al, 2015; Khalil, deClifford, Lam, Subramaniam, 2016; Mergenhagen, Blum, Kugler, et al, 2016) in inpatient settings. The five studies that evaluated the impact of medication reconciliation on reduction of ADEs among all studies identified from the systematic review and our complementary search reported a mean reduction in ADE rates of 74.0% with a range from 42.9% reduction to 90.9% reduction. (Hron, et al, 2015; Mergenhagen, Blum, Kugler, et al, 2016; Boockvar, et al, 2011; Schnipper, et al, 2006) Two of the five studies also quantified the relative reduction in the odds of ADEs in patients exposed to medication reconciliation with odds ratios that ranged from 0.38 to 0.57 (Mergenhagen, Blum, Kugler, et al, 2016; Boockvar, et al, 2011). In addition, to enhance the systematic review by Mueller et al., we searched for studies which showed any evidence of benefit from using medication reconciliation at the time of admission to IPFs. Using Pubmed, we searched a Medical Subject Headings (MeSH) of *medication reconciliation** AND (psychiatr* OR mental) with no publish date requirement. We also manually searched the reference lists of articles identified through this search for additional relevant articles. A total of 24 studies were identified via Mesh term search and 2 studies were considered as relevant via initial independent assessment of titles and subsequent examination of abstracts with or without full-test review. We identified 2 additional relevant articles via the manual literature search, resulting in a total of 4 studies focusing on medication reconciliation in the IPF setting. While none of the studies provided evidence that medication reconciliation had a direct impact on improved patient outcome, they consistently showed that medication reconciliation improved the medication discrepancy detection, demonstrating its potential to reduce ADEs and ultimately improve patient outcomes. According to a multicenter study in three psychiatric facilities, medication discrepancies identified at admission were five times more likely to be associated with potential ADEs than medication discrepancies identified during transitions or at discharge (OR 5.39 95% CI: 2.72 - 10.69) (Keers, et al, 2014). A study conducted in the United Kingdom involving multiple psychiatric hospitals found that discrepancies can include omission of PTA medications (Paton, et al, 2011). Omitted medications such as depot antipsychotics, anti-epileptics, and other medications to treat comorbid medical conditions can cause serious ADEs if not continued during the inpatient stay. Several studies have quantified the number of discrepancies with potential to cause harm. One study found that more than 76% of medication discrepancies in a psychiatric unit were considered potentially harmful at the moderate-to-severe level if medications were not properly reconciled (Brownlie, et al, 2014). Another study of 50 psychiatric inpatients identified that 82% of discrepancies had the potential to cause moderate or severe harm and that 17% of those discrepancies resulted in ADEs including manifestation of affective symptoms, rebound symptoms, progressive psychosis, unmanaged parkinsonism, increased pain, constipation, hypertension, stomach perforation, and nausea. (Prins, Drenth-van Maanen, Kok, Jansen, 2013) Collectively, these findings highlight that medication discrepancies are important targets of medication reconciliation on admission to reduce the rate of ADEs and improve patient safety in the IPF setting. <u>Citations</u> * Byrne SM, Grimes TC, Jago-Byrne MC, Galvin M. Impact of team-versus ward-aligned clinical pharmacy on unintentional medication discrepancies at admission. International journal of clinical pharmacy. 2017;39(1):148-155. doi: 10.1007/s11096-016-0412-4. * Hron JD, Manzi S, Dionne R, et al. Electronic medication reconciliation and medication errors. International journal for quality in health care: journal of the International Society for Quality in Health Care / ISQua. 2015;27(4):314-319. doi: 10.1093/intqhc/mzv046. * Sherr L, Nagra N, Kulubya G, Catalan J, Clucas C, Harding R. HIV infection associated post-traumatic stress disorder and post-traumatic growth--a systematic review. Psychology, health & medicine. 2011;16(5):612-629. doi: 10.1080/13548506.2011.579991. * Gimenez-Manzorro A, Romero-Jimenez RM, Calleja-Hernandez MA, Pla-Mestre R, Munoz-Calero A, Sanjurjo-Saez M. Effectiveness of an electronic tool for medication reconciliation in a general surgery department. International journal of clinical pharmacy. 2015;37(1):159-167. doi: 10.1007/s11096-014-0057-0. * Curatolo N, Gutermann L, Devaquet N, Roy S, Rieutord A. Reducing medication errors at admission: 3 cycles to implement, improve and sustain medication reconciliation. International journal of clinical pharmacy. 2015;37(1):113-120. doi: 10.1007/s11096-014-0047-2. * Leguelinel-Blache G, Arnaud F, Bouvet S, et al. Impact of admission medication reconciliation performed by clinical pharmacists on medication safety. European journal of internal medicine. 2014;25(9):808-814. doi: 10.1016/j.ejim.2014.09.012. * Andreoli L, Alexandra JF, Tesmoingt C, et al. Medication reconciliation: a prospective study in an internal medicine unit. Drugs & aging. 2014;31(5):387-393. doi: 10.1007/s40266-014-0167-3. * Grimes TC, Deasy E, Allen A, et al. Collaborative pharmaceutical care in an Irish hospital: uncontrolled before-after study. BMJ quality & safety. 2014;23(7):574-583. doi: 10.1136/bmjqs-2013-002188. * Lea M, Barstad I, Mathiesen L, Mowe M, Molden E. Effect of teaching and checklist implementation on accuracy of medication history recording at hospital admission. International journal of clinical pharmacy. 2016;38(1):20-24. doi: 10.1007/s11096-015-0218-9. * Becerra-Camargo J, Martinez-Martinez F, Garcia-Jimenez E. A multicentre, double-blind, randomised, controlled, parallel-group study of the effectiveness of a pharmacist-acquired medication history in an emergency department. BMC health services research. 2013;13:337. doi: 10.1186/1472-6963-13-337.

	* Cater SW, Luzum M, Serra AE, et al. A prospective cohort study of medication reconciliation using pharmacy technicians in the emergency department to reduce medication errors among admitted patients. The Journal of emergency medicine. 2015;48(2):230-238. doi: 10.1016/j.jemermed.2014.09.065. * Wang T, Biederman S. Enhance the accuracy of medication histories for the elderly by using an electronic medication checklist. Perspect Health Inf Manag. 2012;9:1-15. * van den Bemt PM, van der Schrieck-de Loos EM, van der Linden C, Theeuwes AM, Pol AG, Dutch CBOWHOHsSG. Effect of medication reconciliation on unintentional medication discrepancies in acute hospital admissions of elderly adults: a multicenter study. Journal of the American Geriatrics Society. 2013;61(8):1262-1268. doi: 10.1111/jgs.12380. * Becerra-Camargo J, Martinez-Martinez F, Garcia-Jimenez E. The effect on potential adverse drug events of a pharmacist-acquired medication history in an emergency department: a multicentre, double-blind, randomised, controlled, parallel-group study. BMC health services research. 2015;15:337. doi: 10.1186/s12913-015-0990-1. * Khalil V, deClifford JM, Lam S, Subramaniam A. Implementation and evaluation of a collaborative clinical pharmacist's medications reconciliation and charting service for admitted medical inpatients in a metropolitan hospital. Journal of clinical pharmacy and therapeutics. 2016;41(6):662-666. doi: 10.1111/jcpt.12442. * Mergenhagen KA, Blum SS, Kugler A, et al. Pharmacist- versus physician-initiated admission medication reconciliation: impact on adverse drug events. The American journal of geriatric pharmacotherapy. 2012;10(4):242-250. doi: 10.1016/j.amjopharm.2012.06.001. * Boockvar KS, Blum S, Kugler A, et al. Effect of admission medication reconciliation on adverse drug events from admission medication changes. Arch Intern Med. 2011;171(9):860-861. doi: 10.1001/archinternmed.2011.163. * Schnipper JL, Kirwin JL, Cotugno MC, et al. Role of pharmacist counseling in preventing adverse drug events after hospitalization. Arch Intern Med. 2006;166(5):565-571. DOI: 10.1001/archinte.166.5.565 * Keers RN, Williams SD, Vattakatuchery JJ, et al. Prevalence, nature and predictors of prescribing errors in mental health hospitals: a prospective multicentre study. BMJ open. 2014;4(9):e006084. doi: 10.1136/bmjopen-2014-006084. 20. Paton C, McIntyre S, Bhatti SF, et al. Medicines Reconciliation on Admission to Inpatient Psychiatric Care: Findings from a UK Quality Improvement Programme. Therapeutic advances in psychopharmacology. 2011;1(4):101-110. doi: 10.1177/2045125311417299. * Brownlie K, Schneider C, Culliford R, et al. Medication reconciliation by a pharmacy technician in a mental health assessment unit. International journal of clinical pharmacy. 2014;36(2):303-309. doi: 10.1007/s11096-013-9875-8. * Prins MC, Drenth-van Maanen AC, Kok RM, Jansen PA. Use of a structured medication history to establish medication use at admission to an old age psychiatric clinic: a prospective observational study. CNS drugs. 2013;27(11):963-969. doi: 10.1007/s40263-013-0103-9.

Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	**The Joint Commission. (2016). *National patient safety goals effective January 1, 2017: Hospital Accreditation Program*. Retrieved on December 13, 2016 from https://www.jointcommission.org/assets/1/6/NPSG_Chapter_HAP_Jan2017.pdf**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	The Joint Commission National Patient Safety Goals for hospitals include the following: Maintain and communicate accurate patient medication information (NPSG.03.06.01). The aspects of this goal that are relevant upon admission to the inpatient setting are stated as: • "Obtain information on the medications the patient is currently taking when he or she is admitted to the hospital or is seen in an outpatient setting. This information is documented in a list or other format that is useful to those who manage medications. Note 1: Current medications include those taken at scheduled times and those taken on an as-needed basis. See the Glossary for a definition of medications. Note 2: It is often difficult to obtain complete information on current medications from a patient. A good faith effort to obtain this information from the patient and/or other sources will be considered as meeting the intent of the EP [element of performance]." • "Compare the medication information the patient brought to the hospital with the medications ordered for the patient by the hospital in order to identify and resolve discrepancies. Note: Discrepancies include omissions, duplications, contraindications, unclear information, and changes. A qualified individual, identified by the hospital, does the comparison."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	None identified
Provide all other grades and definitions from the evidence grading system	Not applicable
Grade assigned to the **recommendation** with definition of the grade	None identified
Provide all other grades and definitions from the recommendation grading system	Not applicable

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	From The Joint Commission: "A panel of widely recognized patient safety experts advise The Joint Commission on the development and updating of NPSGs. This panel, called the Patient Safety Advisory Group, is composed of nurses, physicians, pharmacists, risk managers, clinical engineers and other professionals who have hands-on experience in addressing patient safety issues in a wide variety of health care settings. The Patient Safety Advisory Group works with Joint Commission staff to identify emerging patient safety issues, and advises The Joint Commission on how to address those issues in NPSGs, Sentinel Event Alerts, standards and survey processes, performance measures, educational materials, and Center for Transforming Healthcare projects. Following a solicitation of input from practitioners, provider organizations, purchasers, consumer groups and other stakeholders, The Joint Commission determines the highest priority patient safety issues and how best to address them. The Joint Commission also determines whether a goal is applicable to a specific accreditation program and, if so, tailors the goal to be program-specific." Citation *The Joint Commission. Topic library item. Facts about the National Patient Safety Goals. Development of the Goals. (December 2, 2015). Retrieved November 23, 2016 from https://www.jointcommission.org/facts_about_the_national_patient_safety_goals/
Estimates of benefit and consistency across studies	Not applicable
What harms were identified?	Not applicable
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Not applicable

1a.4 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.

Not applicable

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

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., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.

The Institute for Healthcare Improvement defines medication reconciliation as "the process of creating the most accurate list possible of all medications a patient is taking…and comparing that list against the physician's admission, transfer, and/or discharge orders, with the goal of providing correct medications to the patient at all transition points within the hospital." (Institute for Healthcare Improvement, 2017). While medication reconciliation should occur at all transition points during the inpatient stay, this measure focuses on medication reconciliation on admission because information collected at this transition point is critical to inform treatment decisions during the inpatient stay and at discharge. By collecting adequate information about a patient's PTA medications, recording the information in a single location in the medical record for easy reference, and reconciling this information in a timely manner, clinicians can avoid potentially harmful medication discrepancies. A thorough reconciliation process is important in the IPF setting because pharmacotherapy is a primary form of treatment for patients with severe psychiatric illnesses and the accuracy of self-reported PTA medications may be compromised by severe psychiatric symptoms.

Studies in both the psychiatric and non-psychiatric settings have found that medication discrepancies are present in more than half of medical records for inpatient stays. (Brownlie, 2014; Cornish, 2005). There is evidence to suggest that most medication discrepancies in inpatient medical records result from the failure to collect and reconcile PTA medications. The Multicenter Medication Reconciliation Quality Improvement Study (MARQUIS), which was conducted in six U.S. hospitals, reported an average of 3.35 unintentional medication discrepancies per patient with most medication discrepancies (2.12 per patient) resulting from failure to accurately identify the patient's PTA medications (Salanitro, 2013). The Medications At Transitions and Clinical Handoff (MATCH) study evaluated 651 inpatient stays and found that as many as 85% of admissions with medication errors had errors that originated from incomplete collection of the medication history (Gleason, 2010).

To reduce discrepancies that result from inadequate collection and reconciliation of PTA medications, the Medication Reconciliation on Admission measure is constructed to align with the two elements of performance of The Joint Commission's National Patient Safety Goal (NPSG.03.06.01) on medication safety that are relevant to the admission process (The Joint Commission, 2016). These elements are:

• Obtain information on the medications the patient is currently taking when he or she is admitted to the hospital or is seen in an outpatient setting. This information is documented in a list or other format that is useful to those who manage medications.

• Compare the medication information the patient brought to the hospital with the medications ordered for the patient by the hospital in order to identify and resolve discrepancies.

Citations

* Brownlie, K., Schneider, C., Culliford, R., Fox, C., Boukouvalas, A., Willan, C., & Maidment, I. D. (2014). Medication reconciliation by a pharmacy technician in a mental health assessment unit. Int J Clin Pharm, 36(2), 303-309. doi:10.1007/s11096-013-9875-8

*Cornish, P. L., Knowles, S. R., Marchesano, R., Tam, V., Shadowitz, S., Juurlink, D. N., & Etchells, E. E. (2005). Unintended medication discrepancies at the time of hospital admission. Arch Intern Med, 165(4), 424-429. doi:10.1001/archinte.165.4.424

*Gleason, K. M., McDaniel, M. R., Feinglass, J., Baker, D. W., Lindquist, L., Liss, D., & Noskin, G. A. (2010). Results of the Medications at Transitions and Clinical Handoffs (MATCH) study: an analysis of medication reconciliation errors and risk factors at hospital admission. J Gen Intern Med, 25(5), 441-447. doi:10.1007/s11606-010-1256-6

*Institute for Healthcare Improvement. (2017). Medication reconciliation to prevent adverse drug events. Retrieved from http://www.ihi.org/Topics/ADEsMedicationReconciliation/Pages/default.aspx

*Salanitro, A. H., Kripalani, S., Resnic, J., Mueller, S. K., Wetterneck, T. B., Haynes, K. T., . . . Schnipper, J. L. (2013). Rationale and design of the Multi-center Medication Reconciliation Quality Improvement Study (MARQUIS). BMC Health Serv Res, 13, 230. doi:10.1186/1472-6963-13-230

*The Joint Commission. (2016). National patient safety goals effective January 1, 2017: Hospital Accreditation Program. Retrieved from https://www.jointcommission.org/assets/1/6/NPSG_Chapter_HAP_Jan2017.pdf

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. 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 include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

A sample of nine IPFs from eight states was used to perform the field testing of the measure. Overall measure score results are presented in Table 1b.2a. The average measure score was 50% with a standard deviation of 32% and ranged from 7% to 98% across the nine facilities. Please refer to the NQF Measure Testing Form for all measure testing results.

Table 1b.2a Overall Measure Performance Score

IPF ID // Measure Score (%) // 95% Confidence Interval

IPF 1 // 68 // 59, 77

IPF 2 // 18 // 10, 26

IPF 3 // 77 // 69, 85

IPF 4 // 88 // 82, 94

IPF 5 // 30 // 21,39

IPF 6 // 7 // 2, 12

IPF 7 // 43 // 33, 53

IPF 8 // 98 // 95, 100

IPF 9 // 18 // 10, 26

Average // 50 // N/A

Range // 7–98 // N/A

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.

Not applicable, performance data on the measure are available.

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 maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

Please refer to Tables 1.6-A (Age and Gender) and 1.6-B (Race and Ethnicity) in the NQF Measure Testing Form for the demographic information of the testing population. Because this is a process completed during the admission, we do not anticipate disparities based on sociodemographic status. However, we will monitor for disparities if the measure is implemented.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4

Not applicable because this is a process measure.

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 sub criteria 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):*

De.6. Non-Condition Specific*(check all the areas that apply):*

De.7. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

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.)*

Not available

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)*

No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

S.3.1. For maintenance of endorsement: Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

No

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

Not applicable because this is not a maintenance measure

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) DO NOT include the rationale for the measure.*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Number of patients for whom a designated Prior to Admission (PTA) medication list was generated by referencing one or more external sources of medications and for which all PTA medications have a documented reconciliation action by the end of Day 2 of the hospitalization when the admission date is Day 0.

S.5. 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, time period for data collection, 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 (S.14).

The numerator is operationalized into three key criteria of the medication reconciliation process that must be met:

1. Medications taken by the patient prior to admission are documented on a designated PTA medication list.

2. The PTA medication list is generated using at least one external source to identify the medications taken by the patient prior to admission.

3. All medications listed on the PTA medication list have a reconciliation action to continue, discontinue, or modify by the end of Day 2 of the hospitalization, or if there are no medications on the PTA medication list, the prescriber has signed the document by the end of Day 2 of the hospitalization to indicate his/her review of the PTA medication list.

The first criterion requires that the medical record contain a designated PTA Medication List to document medications that the patient is taking prior to admission. Documenting PTA medications in a designated location eliminates the potential for duplicative or inconsistent documentation of medication histories, avoids the potential for omitted medications, and provides a master source of PTA medication for easy reference by providers. PTA medications may include prescriptions, over-the-counter medications, herbals, vitamin/mineral/dietary (nutritional) supplements, and/or medical marijuana. This criterion aligns with one of the five elements of The Joint Commission's National Patient Safety Goal (NPSG.03.06.01) on medication reconciliation (The Joint Commission, 2016).

The second criterion requires that facilities consult at least one source external to the facility's records to increase comprehensive capture of all active medications on the PTA medication list. Incomplete or inaccurate PTA medication lists may result in inadequate medication reconciliation actions by the prescriber, which may lead to medication errors and ADEs. Given the absence of a single, accurate source of information on PTA medications (gold standard), the measure establishes a minimum standard for compiling PTA medication information rather than being prescriptive regarding which sources should be referenced. This requirement also aligns with other existing NQF-endorsed measures that focus on medication reconciliation. The measure allows for a wide-range of external sources to account for situations where the routinely consulted source fails to generate the information needed. For example, the patient may not be able or willing to provide information on PTA medications or a retail pharmacy may be closed or not willing to disclose PTA medications without obtaining prior patient consent. Therefore, to meet the External Source requirement, the facility can reference one or more of the following sources to compile the PTA medication list:

- Interview of the patient or patient proxy such as a caregiver

- Medication container brought in by patient or patient proxy
- Medication list brought by patient or patient proxy
- Patient support network, such as a group home
- Nursing home
- Outpatient prescriber or emergency department
- Retail pharmacy
- Prescription Drug Monitoring Program (PDMP)
- Electronic prescribing network system (e.g., Allscripts®, Surescripts®) or aggregate pharmacy billing records (such as, claims data using state/federal healthcare plans)

The third and final criterion requires that a licensed prescriber reconciles each medication on the PTA Medication List by the end of Day 2 of the hospitalization and documents whether the medication should be continued, discontinued, or modified. The date of admission is considered Day 0 and subsequent days are considered Day 1 and Day 2 for this measure. If there are no medications on the PTA medication list, the prescriber must sign the document by the end of Day 2 of the hospitalization to indicate his or her review of the PTA medication list for consideration in future treatment decisions. For example, information that indicates the patient is not taking any medications may be important to communicate to the treatment team because there may be a need to initiate treatment of indications that are discovered during admission. Signing the PTA medication list by the end of Day 2 of the hospitalization for patient admissions with no PTA medications also helps to improve communication between members of the care team and other providers during care transitions. To simplify chart abstraction and prevent abstractors from having to distinguish between medications, herbal supplements, and other remedies a patient might take, all entries on the PTA medication list must be reconciled to meet the requirements of the third criterion.

For additional details on each of the data elements included in the measure construct, refer to Appendix A.1, which includes the Data Dictionary and Data Collection Tool.

Citations

*The Joint Commission. (2016). National patient safety goals effective January 1, 2017: Hospital Accreditation Program. Retrieved from https://www.jointcommission.org/assets/1/6/NPSG_Chapter_HAP_Jan2017.pdf

S.6. Denominator Statement *(Brief, narrative description of the target population being measured)*

All patients admitted to an inpatient facility from home or a non-acute setting.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, 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 target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

All adult and pediatric patients admitted to an IPF are eligible to be sampled, regardless of insurance types.

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

The measure applies two exclusion criteria to ensure that it is feasible to complete the medication reconciliation process on admission to the IPF:

1. Patients transferred from an acute care setting

2. Patient admissions with a length of stay less than or equal to 2 days

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, 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.)*

Transfer from an Acute Care Setting:

The first exclusion criterion applies to patient admissions that result from a transfer from an acute care setting, such as another inpatient facility or inpatient unit. This exclusion is applied because medication reconciliation with outpatient medications may have been done at the transferring facility and different medication reconciliation processes are required at the receiving IPF for those admissions to focus on the regimen that was used in the transferring facility. Patient admissions from long-term care facilities and emergency departments are not considered transfers and are included in the denominator for the measure.

Length of Stay Less than or Equal to 2 Days:

The second exclusion criterion applies to patient admissions with lengths of stay shorter than the time needed to adequately complete the medication reconciliation process. The timeframe from admission needed to complete the medication reconciliation process was discussed with the TEP, which recommended a requirement to complete reconciliation by the end of Day 2 if the day of admission is Day 0. They cited instances where patients are admitted on weekends and outpatient providers are not available to ascertain PTA medications or where patients are not stable enough to provide information immediately upon admission. The measure developer also evaluated this timeframe empirically using the field testing data to determine when most facilities could complete the medication reconciliation process. Table 2b2.2 in the NQF Measure Testing Form contains all records with complete medication reconciliation for all medications on the PTA medication list and shows the percentage of those records that had completed the medication reconciliation in one day increments of time from admission. This analysis confirmed the appropriateness of the 2-day timeframe for completing the medication reconciliation process.

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – 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.)*

Not applicable because this measure is not stratified.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

S.13. 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.14. Calculation Algorithm/Measure Logic (*Diagram or 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; time period for data, aggregating data; risk adjustment; etc.)*

To calculate the performance score:

1. Start processing. Run cases that are included in the Initial Patient Population as follows:

a. Find the patients that the performance measure is designed to address (all adult and pediatric patients admitted to the inpatient facility from home or a non-acute setting with a length of stay greater than two days).

2. Check Length of Stay (calculated as the Discharge Date minus the Admission Date).

a. If the Length of Stay is greater 2 days, continue processing and proceed to Transfer From an Acute Care Setting.

b. If the Length of Stay is less than or equal to 2 days, the record will proceed to Measure Category Assignment of B and will not be in the Measure Population. Stop processing.

3. Check Transfer From an Acute Care Setting.

a. If the Transfer From an Acute Care Setting is equal to 1 (Yes), the case was admitted from a transfer from an acute care setting and the record will proceed to Measure Category Assignment of B and will not be in the Measure Population. Stop processing.

b. If the Transfer From an Acute Care Setting is equal to 2 (No), the case was admitted from an admission source other than an acute case setting. Continue processing and proceed to Designated PTA Medication List.

4. Check Designated PTA Medication List.

a. If the Designated PTA Medication List is equal to 1 (Yes), continue processing and proceed to External Source.

b. If the Designated PTA Medication List is equal to 2 (No), the record will proceed to Measure Category Assignment of D and will be in the Measure Population. Stop processing.

5. Check External Source.

a. If External Source is equal to 1 (Yes), continue processing and proceed to Reconciliation Action.

b. If External Source is equal to 2 (No), the record will proceed to Measure Category Assignment of D and will be in the Measure Population. Stop processing.

6. Check Reconciliation Action.

a. If Reconciliation Action is equal to 1 (Yes) or 3 (N/A), continue processing and proceed to Reconciliation Action by End of Day 2.

b. If Reconciliation Action is equal to 2 (No), the record will proceed to Measure Category Assignment of D and will be in the Measure Population. Stop processing.

7. Check Reconciliation Action by the end of Day 2 when the Admission date is Day 0.

a. If Reconciliation Action by End of Day 2 is equal to 1 (Yes), the record will proceed to Measure Category Assignment of E and will be in the Numerator Population. Stop processing.

b. If Reconciliation Action by End of Day 2 is equal to 2 (No), the record will proceed to Measure Category Assignment of D and will be in the Measure Population. Stop processing.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

The measure can use a sample of 100 records or greater. The measure developer will work with CMS to determine the least burdensome sampling approach if the measure is implemented in a program.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

Not applicable

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Paper Medical Records

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

The data dictionary and measure information form that provide instructions for abstracting the data for the measure are included with this application as an attachment. A structured chart abstraction tool with operational data definitions was developed in Microsoft Access for field testing. Prior to implementation, the measure developer will provide a finalized abstraction tool.

S.19. Data Source or Collection Instrument *(available 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. Level of Analysis *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

S.22. <u>COMPOSITE Performance Measure</u> - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

Not applicable because this is not a composite measure.

2. Validity – See attached Measure Testing Submission Form

171004_Med_Rec_nqf_testing_attachment_7.1.docx

2.1 <u>For maintenance of endorsement</u>

Reliability testing: If testing of reliability of the measure score was not presented in prior submission(s), has reliability testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.2 <u>For maintenance of endorsement</u>

Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. Please use the most current version of the testing attachment (v7.1). Include information on all testing conducted (prior testing as well as any new testing); use red font to indicate updated testing.

2.3 <u>For maintenance of endorsement</u>

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes social risk factors is not prohibited at present. Please update sections 1.8, 2a2, 2b1,2b4.3 and 2b5 in the Testing attachment and S.140 and S.11 in the online submission form. NOTE: These sections must be updated even if social risk factors are not included in the risk-adjustment strategy. You MUST use the most current version of the Testing Attachment (v7.1) -- older versions of the form will not have all required questions.

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:**
Measure Title: Medication Reconciliation on Admission
Date of Submission: Click here to enter a date11/1/2017

Type of Measure:

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

Instructions

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

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

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

2b1. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]

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]

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

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

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

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

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

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

Not Applicable. An existing dataset was not available for testing.

1.3. What are the dates of the data used in testing? 1/4/2013 – 8/17/2016

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

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

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*)

A sample of nine Inpatient Psychiatric Facilities (IPFs) from eight states (AZ, CA, CO, LA, MD, MI, WI, and WV) was used to perform the field testing of the measure. Both freestanding facilities and hospital-based units of various sizes and with different types of medical record systems were included in the testing. Table 1.5 provides a breakdown of the characteristics of the IPFs included in the field testing. Each of the nine IPFs were asked to abstract information from 100 patient admissions that met the testing criteria using one of two sampling approaches: (1) selection of the most recent admissions or (2) random selection of admissions. Patient admissions included in the sample had to be from home,

outpatient, emergency, or long-term care. A minimum length of stay of 24 hours was required to be included in the sample because the Measure Developer anticipated that most facilities would need at least 24 hours to adequately complete the medication reconciliation process. The testing sample included adult and pediatric patients and had no restriction on insurance type.

Table 1.5 Field Testing Hospital Characteristics

IPF ID	Location	Type	Bed Size	Data Source
1	West Virginia	Unit	70	EPIC
2	Michigan	Unit	28	McKesson
3	Arizona	Freestanding	90	Paper Medical Records
4	Arizona	Freestanding	75	Paper Medical Records
5	Maryland	Freestanding	322	Allscripts®
6	California	Unit	12	Cerner
7	Louisiana	Unit	38	EPIC
8	Colorado	Freestanding	24	Netsmart TIER® CareRecord™
9	Wisconsin	Freestanding	168	Cerner

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

Testing included a total of 900 patient admissions from the nine IPFs. The testing sample is comprised of adult and pediatric admissions and has no restriction on insurance type. The only inclusion criterion for testing consisted of patient admissions to an inpatient facility from home or a non-acute setting, including admissions from outpatient, emergency department, or long-term care to the IPF for 24 hours or more.

The requirement for admission duration was imposed because the measure required resolution of PTA medication discrepancies with admission orders (i.e., reconciliation action by a licensed prescriber) by the end of Day 2 of the hospitalization when the admission is Day 0.

Tables 1.6-A and 1.6-B show the demographic characteristics of the sample by IPF. IPFs varied notably in the proportion of pediatric and geriatric patients as well as the proportion of African Americans.

Table 1.6-A Age and Gender of Field Testing Population (in percent)

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
No. Records	100	100	100	100	100	100	100	100	100
0-18	0	2	10	45	40	0	4	34	50
19-24	4	20	10	18	12	0	8	5	8
25-34	12	29	28	9	18	0	36	6	7
35-44	18	12	22	11	10	0	23	7	5
45-54	28	19	17	11	9	2	16	12	3
55-64	18	12	11	3	6	5	11	16	2
>65	20	5	1	3	5	93	2	20	25

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
Male	50	59	55	43	55	44	68	51	44

Table 1.6-B Race/Ethnicity of Field Testing Population (in percent)

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
White	96	72	89	89	60	87	40	93	71
Black	3	7	5	4	31	1	57	3	22
Asian/ Pacific Islander	0	0	1	2	3	6	0	1	0
American Indian/ Alaska Native	0	1	5	4	3	1	0	0	1
Other	1	2	0	1	3	5	3	3	0
Unknown Race	0	18	0	0	0	0	0	0	6
Hispanic	1	2	19	24	1	2	3	7	6
Unknown Ethnicity	0	19	0	1	5	55	1	0	4

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.

A 20% random sample of patient records of the originally sampled 100 records for each facility was re-abstracted by a second independent abstractor for the data element reliability analysis (inter-rater reliability).

Facilities were instructed to abstract patient admissions that were 24 hours or more in duration because that was the minimum amount of time deemed to be fair for facilities to complete the medication reconciliation process. However, after testing and reviewing with the TEP, the timeframe for completing the medication reconciliation was extended to 2 days from the admission date. Because dates and times were not collected, the testing sample may include some patient admissions greater than 24 hours but less than 2 days from the admission.

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

Not applicable, measure is not risk adjusted or stratified.

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*)

Data Element Reliability

Two trained abstractors at each IPF independently completed data ascertainment for all measure elements using a random subset of approximately 20 patient records per facility for a total subsample of 175 patient records (Table 2a2.2). There were five cases that could not be used for the inter-rater reliability (IRR) testing because these cases had differing admission dates and/or times and could not be matched to cases reviewed by both abstractors.

Table 2a2.2 Distribution of Records Available for Inter-rater Reliability Analysis Across IPFs

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Total
IRR cases	19	20	18	20	20	20	19	19	20	175

Paired abstractors used a structured medical record abstraction tool developed in Microsoft Excel to independently collect data elements used to define the measure population and to calculate the measure score. Inter-rater reliability between the two abstractors at each site and for each data element used to calculate the measure score was assessed using percent overall agreement and Cohen's Kappa statistic. "Agreed" means the two abstractors provided consistent answers to the same data element question. Cohen's Kappa is a measure of inter-rater agreement that accounts for abstractors' agreement by chance alone. It is standardized on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than by chance (such as, systematic disagreement between abstractors). A common scale is used to interpret Kappa statistics: 0.01–0.20 is considered slight agreement; 0.21–0.40 is fair agreement; 0.41–0.60 is moderate agreement; 0.61–0.80 is substantial agreement; 0.81–0.99 is almost perfect agreement.

To calculate Cohen's Kappa, the abstractors' responses for all patients were organized into four categories (P_{11}: (1, 1), P_{10}: (1, 0), P_{01}: (0, 1) and P_{00}: (0, 0)) for each facility. For each IPF, overall agreement and Cohen's Kappa were calculated.

Cohen's Kappa was calculated based on the following formula:

Cohen's Kappa = $\frac{P_o - P_e}{1 - P_e}$

In which P_o is the observed proportion of agreement and P_e is the expected proportion of agreement.

$P_o = P_{11} + P_{00}$

$P_e = (P_{11} + P_{10}) * (P_{11} + P_{01}) + (P_{00} + P_{10}) * (P_{00} + P_{01})$

Kappa is reported as aggregate across facilities.

Pooled Kappa = $\frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e}$

In which $\bar{P}_o$ is the mean of the P_os and $\bar{P}_e$ is the mean of the P_es across the nine IPFs. The 95% confidence interval of the pooled kappa is $K \pm 1.96 * SE_k$, in which S_e =

$\sqrt{\frac{\bar{P}_o(1-\bar{P}_o)}{n * (1-\bar{p}_e)^2}}$

Performance Measure Score Reliability

The following formula to calculate the reliability of the score for each IPF, reflecting a signal-to-noise ratio.

$$Reliability = \frac{\sigma^2_{Between-IPFs}}{\sigma^2_{Between-IPFs} + \sigma^2_{Within-IPFs}}$$

In which $\sigma^2_{Between-IPFs}$ is the variance of scores between IPFs and $\sigma^2_{Within-IPFs}$ is the variance within IPFs.

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*)

Table 2a2.3-A Percent of Agreement and Cohen's Kappa for Measure Score Data Elements

Data Element	All Records	Agreed	% Agreement	Cohen's Kappa (Confidence Interval)
Designated PTA Medication List	175	166	94.9	0.67 (0.46, 0.88)
External Source	175	131	74.9	0.18 (-0.04, 0.39)
Reconciliation Action	126	118	93.7	0.66 (0.40, 0.91)
Reconciliation Action by End of Day 2 or PTA Medication List signed within 24-hours if no medications	175	157	89.7	0.53 (0.28, 0.77)
Total data elements	651	572	87.9	0.61 (0.53, 0.69)

Table 2a2.3-B Cohen's Kappa within Facilities

	IPF 1	**IPF 2**	**IPF 3**	**IPF 4**	**IPF 5**	**IPF 6**	**IPF 7**	**IPF 8**	**IPF 9**	**Pooled Kappa**
Total data elements (95% CI)	0.65 (0.45, 0.85)	0.82 (0.69, 0.96)	0.18 (-0.14, 0.49)	0.85 (0.56, 1.00)	0.32 (0.11, 0.53)	0.57 (0.39, 0.75)	0.37 (0.10, 0.64)	N/A	1.00 (1.00, 1.00)	0.61 (0.53, 0.69)

Table 2a2.3-C Reliability for Each IPF Final Measure Score

	IPF 1	**IPF 2**	**IPF 3**	**IPF 4**	**IPF 5**	**IPF 6**	**IPF 7**	**IPF 8**	**IPF 9**
Between IPFs σ^2	1143	1143	1143	1143	1143	1143	1143	1143	1143
Within IPF σ^2	21.8	14.8	17.7	10.6	21	6.5	24.5	2.0	14.8
Reliability	0.9813	0.9873	0.9847	0.9908	0.9820	0.9943	0.9790	0.9983	0.9873

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?*)

Data Element Reliability

Inter-rater reliability results are shown in Table 2a2.3-A. For simplicity and computational efficiency, a normal distribution of data was assumed to establish confidence intervals. The confidence intervals for the pooled Cohen's Kappa may therefore generate upper limits smaller than -1.00 or greater than 1.00, which were truncated to -1.00 and 1.00, respectively.

The percentage of overall agreement across the four scoring data elements was 87.9%. The pooled Cohen's Kappa score for the data elements across all nine facilities was 0.61 (95% CI: 0.53, 0.69), indicating substantial agreement (Table 2a2.3-B). The data element with the lowest agreement and Cohen's Kappa score was *External Source* (Criterion 2).

The relatively lower agreement rate for Criterion 2 is likely inherent in current medical record documentation practices, which do not require specification of which sources were used to ascertain PTA medications. Thus, abstractors had to read through admission and progress notes to identify potential sources of PTA medications. It is likely that IPFs will integrate designated fields or check boxes into their medical records if the measure were implemented. This would

simplify and standardize data ascertainment and improve communication with other members of the care team and providers about the source of medications on the PTA medication list.

Table 2a2.3-C provides the Cohen's Kappa score for each facility except IPF 8, which could not be calculated due to perfect agreement between abstractors. Based on the standard interpretation of the scores, IPF 5 had slight agreement, IPF 6 and IPF 7 had moderate agreement, IPF 3 had substantial agreement, and IPF 1, IPF 2, IPF 4, and IPF 9 had perfect agreement. Facility 5 identified several reasons for discrepancies, including how each abstractor handled inconsistent documentation (such as, the reconciliation action was dated after discharge, which was corrected by one abstractor but used verbatim by the other). Discrepancies at IPF 3 and IPF 7 can be explained in part by different interpretations of admission time, which led to different responses to whether some cases completed the medication reconciliation by the end of Day 2 of the hospitalization. Instructions to use the time of the physician admission order have been added to the abstraction instructions to eliminate the need for interpretation and clarify which timestamp should be used.

Performance Measure Score Reliability

The reliability for each IPF measure score is shown in Table 2a2.3-C. The high coefficients reflect small variances within IPF scores and large variance of scores across facilities and indicate that the measure score is highly reliable with a sample of 100 records.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ **Performance measure score**

☐ **Empirical validity testing**

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

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

Systematic Assessment of Face Validity

Face validity of the measure score was obtained by a TEP vote at the conclusion of measure refinement. The TEP was provided with the final measure specifications and presented the results of field testing. After review and discussion during a TEP Meeting, HSAG asked the TEP members to indicate whether they agree, disagree, or are unable to rate the following face validity statement:

> *"The performance scores resulting from the Medication Reconciliation on Admission measure, as specified, can be used to distinguish good from poor facility-level quality related to the process of collecting and reconciling medications on admission to an inpatient facility."*

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

Systematic Assessment of Face Validity

Nineteen of the 21 TEP members voted in agreement that the performance scores resulting from the Medication Reconciliation on Admission measure, as specified, can be used to distinguish good from poor facility-level quality related to the process of collecting and reconciling medications on admission to an inpatient facility. Two TEP members, who did not attend the meeting, did not vote on face validity.

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

Systematic Assessment of Face Validity

The face validity vote (19/19 voting members, 100%) indicates that the measure is viewed as valid by the TEP, which is representative of key stakeholders and experts from the IPF setting.

_____

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section* 2b3

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

The measure applies two exclusion criteria to ensure that it is feasible to complete the medication reconciliation process on admission to the IPF:

3. Patients transferred from an acute care setting
4. Patient admissions with lengths of stay less than or equal to 2 days

Patients transferred from an Acute Care Setting

We were unable to test the transfer exclusion in the field testing sample due to the small sample size. Therefore, we conducted analyses in Medicare fee-for-service (FFS) claims data to estimate the impact of the exclusion by determining the frequency of patients transferred from acute care settings in that subset of the IPF patient population.

Patient Admissions with Lengths of Stay Less than 2 Days

Field testing data were used to empirically evaluate when medication reconciliation actions were completed relative to the day of admission. Table 2b2.2 contains all records with complete medication reconciliation for all medications on the PTA medication list (467 records with a range of 15 to 79 per facility). The table shows the percentage of those records that had completed the medication reconciliation in one day increments of time from admission.

2b2.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Patients Transferred from an Acute Care Setting

Results showed that there were 443,708 Medicare IPF admissions between October 1, 2015, and September 30, 2016. Of these, 26.3% (116,545/443,708) were admissions from transfers from an acute care setting.

Patient Admissions with Lengths of Stay Less than or equal to 2 Days

Table 2b2.2 Percentage of Records with Completed Medication Reconciliation Actions by Day When All PTA Medications in the Record Were Reconciled

Days	IPF 1 n=68	IPF 2 n=65	IPF 3 n=58	IPF 4 n=69	IPF 5 n=49	IPF 6 n=15	IPF 7 n=44	IPF 8 n=79	IPF 9 n=20	Facility Avg	% Across Records	Cumulative % Across Records
Day 0	30.9	7.7	27.6	95.7	4.1	0.0	95.5	25.3	10.0	33.0	37.3	37.3
Day 1	66.2	75.4	25.9	2.9	61.2	26.7	4.6	73.4	55.0	43.5	46.3	83.5
Day 2	2.9	9.2	12.1	0.0	18.4	20.0	0.0	1.3	15.0	8.8	6.6	90.2
Day 3	0.0	0.0	6.9	0.0	4.1	6.7	0.0	0.0	5.0	2.5	1.7	91.9
≥ Day 4	0.0	7.7	27.6	1.5	12.2	0.0	0.0	0.0	15.0	7.1	6.6	98.6
Records with incomplete times	0.0	0.0	0.0	0.0	0.0	46.7	0.0	0.0	0.0	5.2	1.5	100.0

2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Patients Transferred from an Acute Care Setting

Results show that approximately one quarter of IPF patient admissions were admitted from an acute care setting. These results may not be completely generalizable to IPF admissions reimbursed by other insurers or to uninsured patients. However, because the measure will be abstracted using a sample rather than the entire population, the Measure Developer anticipates that the narrower measure population will have a minimal impact on the facility-level denominators. Given that the medication reconciliation process is different for patients transferred from an acute care setting and including these patients could distort measure results, it is appropriate to exclude them from the measure.

Patient Admissions with Lengths of Stay Less than or equal to 2 Days

Results show that 90.2% of records were completely reconciled by the end of Day 2. In other words, if medication reconciliation were completed, turn-around time rarely exceeded three calendar days from admission (or beyond the end of Day 2). Based on these results and the recommendation from the TEP, the measure excludes admissions that are discharged on or before Day 2 of the admission to ensure medication reconciliation can be feasibly completed for the majority of patients.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

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

☐ **No risk adjustment or stratification**

☐ **Statistical risk model with _risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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

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

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

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

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

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

To determine statistically significant differences across the small sample of testing facilities, the Measure Developer calculated the final scores and 95% confidence intervals for each facility using the following formula:

$S_{\text{final score}} = 100 * p$

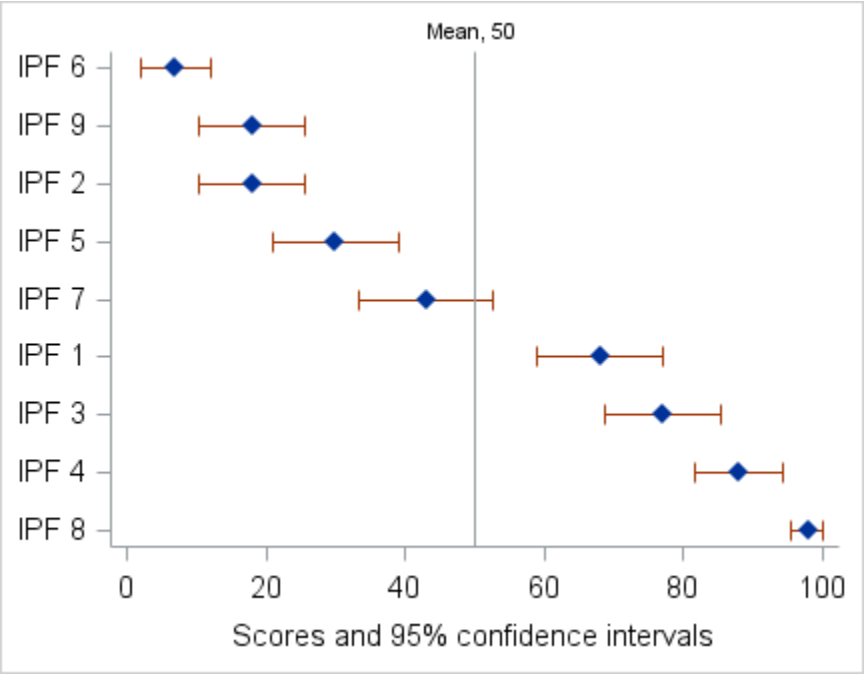
$Se_{\text{final score}} = 100 * \sqrt{\frac{p\,(1-p)}{n}}$, where p represents the proportion of patients meeting all four criteria in the study population and follows a binomial distribution. The 95% confidence interval for the final score is: $S_{\text{final score}} \pm 1.96 * Se_{\text{final score}}$.

Visual examination of a forest plot depicting measure scores and 95% confidence intervals for each facility was used to illustrate whether a given pair of IPFs has statistically significant differences in performance.

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

Figure 2b4.2 displays a forest plot of facility scores with 95% confidence intervals (CI) sorted by score.

Figure 2b4.2 Facility Measure Scores with 95% Confidence Intervals



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

Owing to the broad range of facility-level results, the forest plot illustrates that scores readily discern high- and low-performing facilities. Each facility had scores that were statistically significantly different from at least five of the eight other facilities. All but one facility had scores that were statistically significantly different from the mean measure score. The clinical interpretation of these results suggests substantial differences and ample opportunity for improvement across IPFs in the completeness and timeliness of the medication reconciliation process. This is expected to translate into clinically meaningful differences in reduction of medication discrepancies and preventable ADEs.

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Note that the measure score is largely based on the presence of specific data elements in the medical record. Medication reconciliation within 2 days of admission to an IPF is expected for each patient admission in the measure population. Thus, missing data are considered part of the quality signal in the measure score.

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

Not applicable for the reasons noted in 2b6.1.

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

Not applicable for the reasons noted in 2b6.1.

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), 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*) Update this field for **maintenance of endorsement**.

Some data elements are in defined fields in 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. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

The measure was specified to use manually chart-abstracted data from medical records. This approach was selected for two reasons. First, the setting in which this measure was tested (inpatient psychiatric facilities) primarily used paper records at the time of development. Among IPFs that participate in the IPFQR Program, only about 36% attested to using an EHR system for fiscal year 2016 (CMS, 2016).

This approach was also selected because many of the data elements are not currently collected in structured, computer-readable fields. We anticipate that if this measure were to be implemented, some of the data elements could be collected in structured fields.

Citation:

* Centers for Medicare & Medicaid Services. Inpatient Psychiatric Facility Quality Measure Data – by Facility. 2016. https://data.medicare.gov/Hospital-Compare/Inpatient-Psychiatric-Facility-Quality-Measure-Dat/q9vs-r7wp. Accessed September 13, 2016.

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. Please also complete and attach the NQF Feasibility Score Card.

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. Required for maintenance of endorsement. Describe difficulties (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 instrument-based, consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

If implemented, this measure will rely on abstraction of a sample of medical records by the IPF. After the measure logic and abstraction tool were finalized based on the results of testing, the Measure Developer tested the average time to abstract each record using abstracted data from a total of 36 records from the alpha testing sites. The average time to abstract the eight required data elements was 5.9 minutes. However, we anticipate that this time will go down as facilities update their medication reconciliation processes to collect information, such as the sources referenced to obtain PTA medication information, in more standardized fields in the medical record. For example, a facility could add a list of check boxes for sources referenced rather than collecting that information in notes throughout the record. This would cut down on time to both document the medication reconciliation and to abstract the information to calculate the measure score.

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)*.

There are no fees or other requirements to use this measure as specified.

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.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Not in use	

4a1.1 For each CURRENT use, checked above (update for maintenance of endorsement), provide:

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

Not applicable, this measure is not currently in use.

4a1.2. 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?*)

The measure is currently not publicly reported or in use because this is a new measure.

4a1.3. 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 measure is currently not publicly reported or in use because this is a new measure.

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.

Not applicable, this measure is not currently in use.

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

Not applicable, this measure is not currently in use.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

Not applicable, this measure is not currently in use.

4a2.2.2. Summarize the feedback obtained from those being measured.

Not applicable, this measure is not currently in use.

4a2.2.3. Summarize the feedback obtained from other users

Not applicable, this measure is not currently in use.

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Not applicable, this measure is not currently in use. However, the measure has been revised based on feedback from stakeholders during public comment periods to simplify the measure logic and align data element definitions with similar data elements used by other measures.

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.

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (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.)

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.

The measure is currently not publicly reported or in use because this is a new measure. However, we anticipate that the implementation of this measure will lead to more standardization in the documentation of medication reconciliation at facilities, which will improve communication across providers and may lead to fewer adverse drug events and better patient outcomes. For example, a facility may learn that they are failing the measure frequently at the "one or more external sources of PTA medication" step. They can evaluate whether this is because they are not routinely referencing external sources or they are just not documenting which sources were referenced. If the issue is with the former, they can improve their process for obtaining PTA medications, which may lead to more comprehensive medication lists to inform treatment decisions. If the issue is with the documentation, they can update their medication reconciliation forms to more easily capture the source information, which may reduce re-work in referencing sources that had previously been consulted or lead to more thorough information gathering if members of the team see that sources had not been consulted that they otherwise would have assumed had been referenced.

4b2. 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).

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

Not applicable, this measure has not been implemented yet.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

Not applicable, this measure has not been implemented yet.

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.

Yes

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

0097 : Medication Reconciliation Post-Discharge

0293 : Medication Information

0553 : Care for Older Adults (COA) – Medication Review

0646 : Reconciled Medication List Received by Discharged Patients (Discharges from an Inpatient Facility to Home/Self Care or Any Other Site of Care)

2988 : Medication Reconciliation for Patients Receiving Care at Dialysis Facilities

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

5a. Harmonization of Related Measures

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 harmonized to the extent possible?

Yes

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

The Measure Developer evaluated existing measures in the NQF portfolio to determine whether the Medication Reconciliation on Admission measure would compete with existing measures. Among the five NQF-endorsed measures that evaluate the medication reconciliation process, three (NQF #0097, #0553, #2988) are specified for the outpatient setting and the two (NQF #0293 and #0646) that are specified for the inpatient setting focus on communication of information at discharge. Therefore, the Medication Reconciliation on Admission measure is the only measure that evaluates medication reconciliation on admission to an inpatient facility. To align definitions with other measures that establish a designated timeframe by which a given process must be completed from admission, the Measure Developer harmonized the Medication Reconciliation on Admission measure timeframes with the timeframe specifications of SUB-1 Alcohol Use Screening (NQF 1661) and TOB-1 Tobacco Use Screening (NQF 1651), developed by The Joint Commission. Both measures define the length of stay in calendar days. Standardizing definitions for calculating length of stay using the admission and discharge dates without factoring-in the admission and discharge times will not only help reduce confusion across measures but also help to improve the reliability of the measure scores by eliminating the need to capture times, which were found to be unreliable during field testing. To develop the three data elements associated with the medication reconciliation process, the Measure Developer compared the conceptual descriptions and definitions of five NQF-endorsed measures (NQF 0553, NQF 2988, NQF 0293, NQF 0646, and NQF 0097) that evaluate the medication reconciliation process. Four of the five measures explicitly require a designated medication list. For this measure, the Measure Developer operationalized that requirement with the Designated PTA Medication List data element. Of the three measures that required collection of medications, two had requirements for the types of sources that should be referenced to compile the list. For the Medication Reconciliation on Admission measure, the Measure Developer set to establish a minimum standard and aligned with the approach to require “one or more external sources.” While several measures required the type of information to be collected on each medication, the Measure Developer decided not to include those data elements in this measure given the high performance and low variation for those data elements in testing. Each of the measures defines the process of reconciling the medications on the list differently. The Measure Developer incorporated aspects of each definition that are most applicable to the IPF setting. For example, the Measure Developer aligned with measures that require that the reconciliation be completed by a prescriber and that there be documentation of whether each medication be continued, modified, or discontinued. Finally, the Measure Developer considered different approaches to scoring the measure. Four of the five NQF-endorsed measures require that all aspects of the medication reconciliation process be completed for a patient to pass the measure. The fifth measure evaluates the number of patient months for which the medication reconciliations were completed, however, this is only applicable in the outpatient setting. Therefore, the Measure Developer aligned the scoring approach to produce measure scores that represent the percentage of patient admissions that meet all the medication reconciliation criteria.

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.)
This measure complements other existing measures because it focuses on the completion of the medication reconciliation process by the end of Day 2 of the hospitalization to the facility, which is not addressed by any existing measure. Medication reconciliation on admission is important to inform accurate medication reconciliation at discharge, which is evaluated by two of the existing measures. Medication reconciliation on admission also ensures that efforts to reconcile medications in the outpatient setting are continued at the transition to the inpatient setting.

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: Appendix_A.1_Supplemental_Materials_Med_Rec.docx

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Vinitha, Meyyur, vinitha.meyyur@cms.hhs.gov, 410-786-8819-

Co.3 Measure Developer if different from Measure Steward: Health Services Advisory Group, Inc. (HSAG)

Co.4 Point of Contact: Megan, Keenan, mkeenan@hsag.com, 616-425-1997-

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.

Medication Reconciliation Workgroup Members:

Workgroup Members from the TEP

- Kathleen Delaney, PhD, PMH-NP, RN
- Jonathan Delman, PhD, JD, MPH
- Irene Ortiz, MD, MSW
- Elvira Ryan, MBA, BSN, RN
- Lisa Shea, MD

Workgroup Members from the University of Florida

- Regina Bussing, MD

- Michael Shapiro, MD
- Ben Staley, PharmD, BCPS
- Gigi Lipori, MBA
- Carl Henricksen, MS
- Xinyue Liu, Ph.D
- Nakyung Jeon, MPH, PhD
- Almut Winterstein, RPh, PhD, FISPE

Inpatient Psychiatric Facility Measure Development and Maintenance TEP 2014-2016

Alisa Busch, MD, MS

Director, Integration of Clinical Measurement & Health Services Research

Chief, Health Services Research Division, Partners Psychiatry and Mental Health

Assistant Professor of Psychiatry and Health Policy, Harvard Medical School

Kathleen Delaney, PhD, PMH-NP, RN

Professor, Rush College of Nursing

Jonathan Delman, PhD, JD, MPH

Assistant Research Professor, Systems and Psychosocial Advances Research Center, University of Massachusetts Medical School

Frank Ghinassi, PhD, ABPP

President and CEO, Rutgers University Behavioral Health Care

Adjunct Associate Professor of Psychiatry, University of Pittsburgh School of Medicine

Eric Goplerud, PhD

Senior Vice President, Director of Public Health Department, NORC at the University of Chicago

Geetha Jayaram, MD

Associate Professor, Schools of Medicine, Health Policy and Management and the Armstrong Institute for Patient Safety, Johns Hopkins University

Charlotte Kauffman, MA, LCPC

Service Systems Coordinator, State of Illinois-Division of Mental Health

Tracy Lenzini, BS

Executive Director, Grand Traverse Health Advocates

Kathleen McCann, RN, PhD

Director of Quality and Regulatory Affairs, National Association of Psychiatric Health Systems

Gayle Olano-Hurt, MPH, CPHQ, PMC

Director Data Management, Outcomes Measurement & Research Administration, Sheppard Pratt Health System

Mark Olfson, MD, MPH

Professor of Psychiatry, Columbia University Medical Center Department of Psychiatry; New York State Psychiatric Institute

Irene Ortiz, MD, MSW

Medical Director, Molina Healthcare of New Mexico

Thomas Penders, MS, MD, DLFAPA

Medical Director, Inpatient Psychiatry, Vident Medical Center

Associate Professor, Brody School of Medicine Department of Psychiatry, East Carolina University

Elvira Ryan, MBA, BSN, RN

Associate Project Director, Division of Healthcare Quality Evaluation, The Joint Commission

Lucille Schacht, PhD

Senior Director, Performance and Quality Improvement, National Association of State Mental Health Program Directors Research Institute, Inc.

Lisa Shea, MD

Medical Director, Butler Hospital

Thomedi Ventura, MS, MSPH

Program Evaluator, Telligen

Inpatient Psychiatric Facility Measure Development and Maintenance TEP 2016-2018

Robert Cotes, MD

Medical Director, Inpatient Psychiatry at Grady Memorial Hospital

Kathleen Delaney, PhD, PMH-NP, FAAN

Professor, Rush College of Nursing

Vikas Duvvuri, MD, PhD

Medical Director, Fremont Hospital

Nola Harrison, ACSW, LSCW, LSW-A

Director, St. Anthony Hospital

Nora Lott Haynes, Med, EdS

Coordinator, NIMH Research Project, NAMI Savannah

Gayle Olano Hurt, MPA, CPHQ, PMC

Director Data Management, Outcomes Measurement, and Research Administration, Sheppard Pratt Health System

Mary Jane Krebs, FACHE

President, Spring Harbor Hospital

Kathleen McCann, RN, PhD

Director of Quality and Regulatory Affairs, National Association of Psychiatric Health Systems

Marsden McGuire, MD, MBA

Deputy Chief Consultant, Mental Health Services, Department of Veterans Affairs

Margaret Paccione-Dyszlewski, PhD

Director of Clinical Innovation, Bradley Hospital

Michael Peterson, MD, PhD

Director of Hospital Psychiatric Services, University Hospital

Nancy Purtell , MBA/HCM, RN

Assistant Vice President, Behavioral Health Services, Hospital Corporation of America (HCA)

Jessica Ross, MD, MS

Assistant Clinical Professor; Chief Informatics Officers, UCSF and Zuckerberg SF General Hospital, Department of Psychiatry

Elvira Ryan, MBA, BSN, RN

Associate Project Director, The Joint Commission

Lisa Shea, MD

Medical Director, Butler Hospital

Mary Kay Shibley, MSN, RN

Clinical Informaticist, Sharp Mesa Vista Hospital

Ann M. Sissler, MSW, LSW, ACSW

Senior Director, Quality and Patient Safety, Behavioral Health Services, Westchester Medical Center

Johan Smith, MBA

Vice President of Health Informatics, Universal Health Services, Horizon Health, Mental Health Outcomes

Julia Sullivan, MSN, RN-BC

Assistant Professor, Nursing, Santa Fe College

James D. Tew, Jr., MD

Medical Director, Quality and Clinical Pathways, Western Psychiatric Institute & Clinic

Michael Trangle, MD

Senior Medical Director, HealthPartners/Regions Hospital

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? Not applicable, this is a new measure.

Ad.5 When is the next scheduled review/update for this measure?

Ad.6 Copyright statement: Not applicable; the measure in in the public domain.

Ad.7 Disclaimers: None

Ad.8 Additional Information/Comments: The original version of the Medication Reconciliation on Admission measure (NQF #3207) was constructed as a composite measure. The measure was reviewed by the NQF Behavioral Health Standing Committee (BHSC) in February 2017. The composite measure was not recommended for endorsement by the NQF BHSC because the committee vote indicated that the measure did not pass the evidence criterion, which is required for further evaluation.

The measure was subsequently re-specified from a composite measure into a process of care measure. Additionally, the evidence supporting the measure was enhanced and the measure algorithm was simplified to address the concerns raised by the NQF BHSC committee. This revised version of the Medication Reconciliation on Admission measure is being submitted to NQF as a new measure for endorsement consideration.