MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0104e
Measure Title: Adult Major Depressive Disorder (MDD): Suicide Risk Assessment
Measure Steward: PCPI
Brief Description of Measure: Percentage of patients aged 18 years and older with a diagnosis of major depressive disorder (MDD) with a suicide risk assessment completed during the visit in which a new diagnosis or recurrent episode was identified
Developer Rationale: This measure aims to improve rates of clinician assessment of suicide risk during an encounter where a new or recurrent episode of major depressive disorder is identified. In an epidemiologic study (2010) of mental illness in the United States with a large, representative sample, 69% of respondents with lifetime suicide attempts had also met diagnostic criteria for major depressive disorder. When considering other mood disorders related to depression, such as dysthymia and bipolar disorders, this rate increases to 74%. (1) In a 2014 study conducted by Ahmedani et al, 50% of individuals who completed a suicide had been seen in a health care setting within four weeks prior. (2) Better assessment and identification of suicide risk in the health care setting should lead to improved connection to treatment and reduction in suicide attempts and deaths by suicide.

(1) Bolton, J. M., & Robinson, J. (2010). Population-Attributable Fractions of Axis I and Axis II Mental Disorders for Suicide Attempts: Findings From a Representative Sample of the Adult, Noninstitutionalized US Population. American Journal of Public Health, 100(12), 2473–2480. doi:10.2105/ajph.2010.192252

(2) Ahmedani, B. K., Simon, G. E., Stewart, C., Beck, A., Waitzfelder, B. E., Rossom, R., … Solberg, L. I. (2014). Health Care Contacts in the Year Before Suicide Death. Journal of General Internal Medicine, 29(6), 870–877. doi:10.1007/s11606-014-2767-3

Numerator Statement: Patients with a suicide risk assessment completed during the visit in which a new diagnosis or recurrent episode was identified
Denominator Statement: All patients aged 18 years and older with a diagnosis of major depressive disorder (MDD)
Denominator Exclusions: None

Measure Type: Process
Data Source: Electronic Health Records
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Feb 28, 2014

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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

- Evidence supporting the measure includes the American Psychiatric Association (APA) Practice guideline for the treatment of patients with major depressive disorder. Oct 2010. Reaffirmed Oct 2015.
    - 1170 articles are cited in 2010 version
    - 773 additional articles were reviewed for 2015 reaffirmation.
    - The guideline has a Category I recommendation which indicated substantial clinical confidence.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: 2015 reaffirmation of the guideline.

Exception to evidence: N/A

Questions for the Committee:

- *The evidence provided by the developer has been updated, but is directionally the same as the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

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) → high (Box 5b) → high

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 2015 CMS Physician Quality Reporting Initiative data from the previous submission demonstrates a gap in care. In 2015 the average performance rate was 71.3%. The developer showed performance rates from 2012-2015 that demonstrated a range of 71.3% to 86% during this time frame of providers who document the presence or absence of suicidal ideation and who assess for suicide risk.

Disparities

- The developer was unable to provide data on disparities from the CMS Physician Quality Reporting Initiative, nor identify studies that examined disparities in suicide assessment rates. The developer included findings in suicide disparities from the CDC's 2017 report: Suicide Trends Among and Within Urbanization Levels by Sex, Race/Ethnicity, Age Group, and Mechanism of Death- United States, 2001-2015.

Questions for the Committee:

- *There was no data on disparities in suicide assessment rates provided, 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:

**Research supports conducting a suicide risk assessment for those diagnosed with MDD. Asking about suicide is often ignored. This is a process measure and would be a significant contribution to the field. I am concerned that the intent of the measure is only to assess at time of first diagnosis or recurrence (after 105 days). May want to consider including additional risk factors that would highlight need for suicide risk assessment should the individual's circumstances change.
**Substantial evidence exists to support a measure related to suicide assessment of individuals with major depressive disorder.
**Evidence has not changed. I do not think we need to review it again and could move forward.
**The evidence and importance for this process measure is high.
**Process measure.

1b. Performance Gap

Comments:

**The researchers show a gap in care such that patients are frequently seen in health care settings in the months leading up to a death by suicide and these are missed opportunities to screen for suicide. Furthermore, patients with known mood disorders are at increased risk for suicide and therefore applying this measure to those recently diagnosed with MDD makes sense. I think this measure should be applied to all settings where the patient is diagnosed with MDD or a recurrence of MDD. The measure states emergency departments, outpatient, behavioral health day treatment. I'd like to ensure that it includes all health care settings where a diagnosis of MDD is initiated.
**There is a lot of information on disparities in completed suicides, but I am not aware of any studies currently conducted on disparities in risk assessment of suicide. That said, the developers provide sufficient evidence of a performance gap in this area.
**There is a gap - 71% or 86% comply. Better than other measures but being suicide any gap might be a problem. No disparities from claims seems strange as the Center for Suicide Prevention does show disparities by race and by age https://www.sprc.org/racial-ethnic-disparities
**There is adequate evidence of both a gap and some improvement over time.
**A performance gap exists--average performance rate of 71.3%.

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**
Evaluators: NQF Staff

Evaluation of Reliability and Validity: Link A

Questions for the Committee regarding reliability:

- *The 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 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**

Committee pre-evaluation comments

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

2a1. Reliability – Specifications

Comments:

**Data elements are clearly defined.
**High.

2a2. Reliability – Testing

Comments:

**My concern is without the use of a standardized tool, it will be unclear how physicians are assessing suicide risk. For example, those who say "you aren't having thoughts of killing yourself, right?" It is possible results will be an underestimate of the prevalence of suicide should they ask in this way. While I would recommend a standardized risk assessment, this is a great step towards better identification of those at risk for suicide.
**No. Testing shows high reliability at .97
**Not sure if this comes under Reliability or Usability. The measure EXCLUDES telehealth - seems unacceptable to me given we are doing an increased volume of behavioral health evaluations and treatment via telehealth.
**Adequate.
**No--reliability is high

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

Comments:

**No concerns.
**Can it be valid excluding telehealth?
**I find it interesting that the design of this measure is based upon searching the EHR and automatically harvesting the data. In reality the yield from this has been low and the majority of the positive results have been obtained from manual review of the EHR. Presumably over time, EHRs will incorporate more analyzable field and get better. Currently there is some risk to validity that systems with less resources and limited ability to perform manual reviews, will perform at a significantly lower level. in reality.
**Moderate validity.

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

Comments:

**Yes Any patient receiving services via telehealth.
**No other problems.

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 data elements are routinely generated and used during the provision of care.
- Data element feasibility scorecard was calculated across three EHR vendors (Epic, NextGen, and Point Click Care), all elements are in a structured format in the EHRs with the exception of ED visit was found to be not defined in two EHRs.
- Identifying patients to meet numerator may be challenging as suicide risk assessment is consistently documented in free text notes requiring manual review.

Questions for the Committee:

- *Does the committee have any concerns about the feasibility of identifying patients for the numerator?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**Concern was already articulated by developers regarding difficulty in extrapolating this information from the medical record but this shouldn't be a reason to not approve this measure.
**Data can be extracted electronically; however, it must be extracted manually because it is not readily available in a structured format, which seems like a feasibility issue.
**Telehealth issue.
**Adequate.
**Feasibility moderately good.

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**

Accountability program details

- The measure is used in the CMS' Merit-based Incentive Payment System (MIPS). Prior to 2016, it was used in the Physician Quality Reporting System (PQRS).
- The developer notes that CMS intends to "make all measures under MIPS quality performance category available for public reporting on Physician Compare in the transition year of the Quality Payment Program, as technically feasible." 2018 data for this measure will be available for public reporting on Physician Compare in late 2019.

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

- Feedback was obtained by cross-specialty, multi-disciplinary work groups during the measure development process.
- Measure developer (PCPI) obtains feedback via a public comment period via an online survey tool as well as solicits implementer feedback.
- Measure developer reports feedback on measure that suggested the initial suicide risk assessment was too complex, and has in response to this feedback, reduced the number of suicide risk assessment components

to the four most essential ones. In addition, the measure logic has been modified to include a lookback period for a prior diagnosis of new or recurrent MDD.

Additional Feedback: 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

- While the PCPI creates measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible.

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

Potential harms

- None reported

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:

**It is great that developers took the feedback of measure implementers and modified the risk assessment down to 4 main items. It would be helpful to know what physicians will do for those who screen positive for suicide and what kind of guidance will be offered thru use of this measure.
**Currently used in the Merit-based Incentive Payment System (MIPS).
**Adequate.

4b1. Usability – Improvement
Comments:
**The harm will be if physicians do not follow up on the suicide risk endorsed by patients. But assessing for suicide risk in those with MDD is a good first step to ensure that more patients are identified. Also, for physicians (and non mental health providers) who don't use a standardized tool, scoring, and thus next steps, aren't clear. My concern is physicians taking seriously the concerns that patients under their care and not relying on prescriptions for those at risk, rather than a brief evidenced-based intervention.
**Consistent suicide risk assessment can lead to proper identification and treatment, leading to reductions in attempts and completed suicides. This is a common sense measure for the population of focus.
**I am not sure that this is that useful as it will miss a significant number of patients. As identified in the Brief Description:
(1) In a 2014 study conducted by Ahmedani et al, 50% of individuals who completed a suicide had been seen in a health care setting within four weeks prior.

If they only look at "new" episodes then we are missing all of the routine ongoing MDD treatment and making sure they also get a suicide eval. Major concerns here for me.
**Adequate.
**Data is currently publically available. Used in MIPS.

Criterion 5: Related and Competing Measures

Related or competing measures

There are no competing measures. The developer notes the following related measure:

- NQF 1365: Child and Adolescent Major Depressive Disorder (MDD) Suicide Risk Assessment

Harmonization

- Measure 1365 and 0104 were both developed by PCPI and harmonized to the extent possible.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 0104e
Measure Title: Adult Major Depressive Disorder (MDD): Suicide Risk Assessment

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document
   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?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐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, skip Questions #3-8, then go to Question #9)

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

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #5)

   The developer used a beta-binomial model to assess the signal to noise ratio. The overall average reliability is 0.94. The reliability above the minimum level of quality reporting events is 0.97.

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #6)

   ☐Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

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

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   REFERENCE: Testing attachment, section 2a2.2
   ***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 #8)
   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   ***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*
   ☒Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

***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 #13)

☐No (go to Question #13)

☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)
Developer does not indicate missing data.

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☐Yes (go to Question #19)
☒No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***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 #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☐Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☒Yes (go to Question #22)

☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

REFERENCE: *Testing attachment, section 2b1.*

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 #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

Conducted correlation analysis with Depression Utilization of the PHQ-9 Tool (PQRS #371) – hypothesis that there is a positive association between patients with major depressive disorder that receive a suicide risk assessment and those that have had a PHQ-9 administered.

23. **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 (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

A positive correlation was found between the measures with a coefficient of 0.39 and p-value equals 0.45. Due to small sample size (n = 120), the correlation did not reach statistical significance.

24. 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

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 #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0104
Measure Title: Major Depressive Disorder (MDD): Suicide Risk Assessment
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: 4/2/2018

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: Suicide risk assessment

☐ 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.

Process>>> Suicide Risk Assessment>>> physician adherence to guideline recommendations>>> accurate identification of suicide risk/suicidal intent>>> appropriate treatment, reduction in patient risk/suicide attempts/death

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

****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**	American Psychiatric Association (APA). Practice guideline for the treatment of patients with major depressive disorder. 3rd ed. Arlington (VA): American Psychiatric Association (APA); 2010 Oct. 152 p. Reaffirmed Oct 2015.

• **Date** • **Citation, including page number** • **URL**	https://www.psychiatry.org/psychiatrists/practice/clinical-practice-guidelines
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 careful and ongoing evaluation of suicide risk is necessary for all patients with major depressive disorder [I]. (APA, 2010, p. 15) Such an assessment includes specific inquiry about suicidal thoughts, intent, plans, means, and behaviors; identification of specific psychiatric symptoms (e.g., psychosis, severe anxiety, substance use) or general medical conditions that may increase the likelihood of acting on suicidal ideas; assessment of past and, particularly, recent suicidal behavior; delineation of current stressors and potential protective factors (e.g., positive reasons for living, strong social support); and identification of any family history of suicide or mental illness [I]. (APA, 2010, p. 15)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	None
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	APA Guideline: Category I
Provide all other grades and definitions from the recommendation grading system	Each recommendation is identified as falling into one of three categories of endorsement, indicated by a bracketed Roman numeral following the statement. The three categories represent varying levels of clinical confidence: [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 description of the evidence review in the APA guideline did not address the overall quantity of studies in the body of evidence related to performing suicide risk assessment. However, 1170 articles are cited in the guideline´s reference section. An additional 773 articles were reviewed for the 2015 reaffirmation of guideline currency. The quality of the body of evidence supporting the measure focus was not addressed in the APA guideline. However, the following paragraph was included: This document represents a synthesis of current scientific knowledge and rational clinical practice regarding the treatment of patients with major depressive disorder. It strives to be as free as possible of bias toward any theoretical approach to treatment. In order for the reader to appreciate the evidence base behind the guideline recommendations and the weight that should be given to each recommendation, the summary of treatment recommendations is keyed according to the level of confidence with which each recommendation is made. Each

	rating of clinical confidence considers the strength of the available evidence. When evidence from randomized controlled trials and meta-analyses is limited, the level of confidence may also incorporate other clinical trials and case reports as well as clinical consensus with regard to a particular clinical decision. In the listing of cited references, each reference is followed by a letter code in brackets that indicates the nature of the supporting evidence. (APA, 2010)
Estimates of benefit and consistency across studies	The consistency of results across studies supporting the measure focus was not addressed in the APA guideline. However, the relevant APA recommendation statement received a Category I recommendation which indicates that the practice was recommended with substantial clinical confidence.
What harms were identified?	No harms were identified.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The American Psychiatric Association reaffirmed the currency of the guideline in October 2015. An additional 773 articles were reviewed for this reaffirmation. The review of these additional articles did not impact the recommendations supporting the focus of this measure. An additional review of studies examining screening for suicide risk in patients with depression published after October 2015 did not turn out any findings that would change the focus of this measure.

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.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0104e
Corresponding Measures:
De.2. **Measure Title:** Adult Major Depressive Disorder (MDD): Suicide Risk Assessment
Co.1.1. **Measure Steward:** PCPI
De.3. **Brief Description of Measure:** Percentage of patients aged 18 years and older with a diagnosis of major depressive disorder (MDD) with a suicide risk assessment completed during the visit in which a new diagnosis or recurrent episode was identified
1b.1. **Developer Rationale:** This measure aims to improve rates of clinician assessment of suicide risk during an encounter where a new or recurrent episode of major depressive disorder is identified. In an epidemiologic study (2010) of mental illness in the United States with a large, representative sample, 69% of respondents with lifetime suicide attempts had also met diagnostic criteria for major depressive disorder. When considering other mood disorders related to depression, such as dysthymia and bipolar disorders, this rate increases to 74%. (1) In a 2014 study conducted by Ahmedani et al, 50% of individuals who completed a suicide had been seen in a health care setting within four weeks prior. (2) Better assessment and identification of suicide risk in the health care setting should lead to improved connection to treatment and reduction in suicide attempts and deaths by suicide.

(1) Bolton, J. M., & Robinson, J. (2010). Population-Attributable Fractions of Axis I and Axis II Mental Disorders for Suicide Attempts: Findings From a Representative Sample of the Adult, Noninstitutionalized US Population. American Journal of Public Health, 100(12), 2473–2480. doi:10.2105/ajph.2010.192252

(2) Ahmedani, B. K., Simon, G. E., Stewart, C., Beck, A., Waitzfelder, B. E., Rossom, R., … Solberg, L. I. (2014). Health Care Contacts in the Year Before Suicide Death. Journal of General Internal Medicine, 29(6), 870–877. doi:10.1007/s11606-014-2767-3

S.4. **Numerator Statement:** Patients with a suicide risk assessment completed during the visit in which a new diagnosis or recurrent episode was identified
S.6. **Denominator Statement:** All patients aged 18 years and older with a diagnosis of major depressive disorder (MDD)
S.8. **Denominator Exclusions:** None

De.1. **Measure Type:** Process
S.17. **Data Source:** Electronic Health Records
S.20. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Feb 28, 2014

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

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

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

0104_nqf_evidence_attachment_7.1.docx

1a.1 <u>For Maintenance of Endorsement:</u> 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

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)*

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

This measure aims to improve rates of clinician assessment of suicide risk during an encounter where a new or recurrent episode of major depressive disorder is identified. In an epidemiologic study (2010) of mental illness in the United States with a large, representative sample, 69% of respondents with lifetime suicide attempts had also met diagnostic criteria for major depressive disorder. When considering other mood disorders related to depression, such as dysthymia and bipolar disorders, this rate increases to 74%. (1) In a 2014 study conducted by Ahmedani et al, 50% of individuals who completed a suicide had been seen in a health care setting within four weeks prior. (2) Better assessment and identification of suicide risk in the health care setting should lead to improved connection to treatment and reduction in suicide attempts and deaths by suicide.

(1) Bolton, J. M., & Robinson, J. (2010). Population-Attributable Fractions of Axis I and Axis II Mental Disorders for Suicide Attempts: Findings From a Representative Sample of the Adult, Noninstitutionalized US Population. American Journal of Public Health, 100(12), 2473–2480. doi:10.2105/ajph.2010.192252

(2) Ahmedani, B. K., Simon, G. E., Stewart, C., Beck, A., Waitzfelder, B. E., Rossom, R., … Solberg, L. I. (2014). Health Care Contacts in the Year Before Suicide Death. Journal of General Internal Medicine, 29(6), 870–877. doi:10.1007/s11606-014-2767-3

1b.2. **Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis**. *(<u>This is required for maintenance of endorsement</u>. 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.*

Hepner and colleagues (2007) found that primary care physicians (PCPs) assess for suicide only 24% of the time in patients with depression.(1) In the same study, only 28% of PCPs adhered to the quality indicator "Treatment for suicidal ideation among patients not already followed in mental health care."(1) McGlynn and colleagues (2003) found that only 25.8% of PCPs document the presence or absence of suicidal ideation during the first or second diagnostic visit.(2) The same study showed that only 28.9% of patients who have suicidality and have any of the following risk factors: psychosis, current alcohol or drug abuse or dependency, and specific plans to carry out suicide (eg, obtaining a weapon, putting affairs in order, making a suicide note) are hospitalized.(2) Additionally, Luoma and colleagues (2002) found that 40% of patients who completed suicide had seen their primary care physician in the past month.(3)

2015 Physician Quality Reporting System (PQRS) Experience Report

2015 is the most recent year for which PQRS Experience Report measure data are available. The average performance rates on Adult Major Depressive Disorder (MDD): Suicide Risk Assessment over the last several years are as follows:

- 2012: 77.0%
- 2013: 76.7%

- 2014: 86.0%
- 2015: 71.3%

2015 Reporting Experience, Including Trends (2007-2016), Physician Quality Reporting System. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/AnalysisAndPayment.html

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. As a result, performance rates may not be nationally representative. Beginning in 2015, the program imposed payment penalties for non-participants based on 2013 performance

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

Performance variability for suicide assessment in MDD is well demonstrated in clinical quality literature. In a 2012 study that examined provider intent to assess for suicidality in patients with MDD, 404 Primary Care Providers (PCPs) were shown a standardized virtual patient. 98% of the physicians accurately diagnosed that patient with depression. However, only 36% reported a recommendation to assess for suicide risk. Statistically significant variation also existed in provider demographics between assessors and non-assessors, suggesting inconsistent application of suicide assessment guidelines in patients with MDD. (1) In another study (2011) featuring primary care patients with positive depression screens, suicide-related discussion occurred in only 11% of encounters. (2) Finally, in their study that included 281 depression-related visits, McGlynn and colleagues (2003) found that only 25.8% of PCPs document the presence or absence of suicidal ideation during the first or second diagnostic visit. (3)

(1) Hooper, L. M., Epstein, S. A., Weinfurt, K. P., DeCoster, J., Qu, L., & Hannah, N. J. (2012). Predictors of Primary Care Physicians' Self-reported Intention to Conduct Suicide Risk Assessments. The Journal of Behavioral Health Services & Research, 39(2), 103–115. doi:10.1007/s11414-011-9268-5
(2) Vannoy, S. D., & Robins, L. S. (2011). Suicide-related discussions with depressed primary care patients in the USA: gender and quality gaps. A mixed methods analysis. BMJ Open, 1(2), e000198–e000198. doi:10.1136/bmjopen-2011-000198
(3) McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA. The Quality of Health Care Delivered to Adults in the United States. N Engl J Med 2003;348:2635-2645.

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.*

While this measure is included in several federal reporting programs, those programs have not yet made disparities data available for us to analyze and report.

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

We were not able to identify any studies that examined disparities in suicide assessment rates among people with MDD. However, several well-established disparities exist among individuals who complete a suicide.

Key findings in suicide disparities from the CDC's 2017 Report: Suicide Trends Among and Within Urbanization Levels by Sex, Race/Ethnicity, Age Group, and Mechanism of Death—United States, 2001-2015.

- Suicide was the 10th leading cause of death in 2015, with a total count of 44,193 deaths.
- The age adjusted suicide rate increased 21.6% during 2001-2015.
- Suicide rates are higher for males than for females.
- Suicide rates are higher for adults aged >=45 than for adolescents and young adults.

- Overall suicide rates are higher for non-Hispanic whites and American Indian/Alaskan Native populations than other ethnic groups.
- Suicide rates by sex, race/ethnicity, age group, and mechanism of death are higher in rural communities than urban ones.

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):*
Behavioral Health, Behavioral Health : Depression, Behavioral Health : Suicide

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):*
Elderly

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 specifications are included as an attachment with this submission. Additional measure details may be found at: eCQI Resource Center webpage https://ecqi.healthit.gov/eligible-professional-eligible-clinician-ecqms . Value set details at VSAC we

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:** EP_EC_CMS161v6_NQF0104_MDD_SuicideRisk.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:** 0104_MDD_SuicideRisk_ValueSets_2017September29.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.

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.
Yes

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.

This measure is intended to only require a suicide risk assessment at the visit in which a new or recurrent episode of Major Depressive Disorder (MDD) is diagnosed. Measure implementers have given us feedback that identifying the visit in which a new or recurrent episode of Major Depressive Disorder (MDD) is diagnosed has been challenging, as the measure logic had been indicating every visit for MDD as a new recurrent episode of MDD. After discussion, the clinical experts agreed that the initial population logic should be modified and to introduce a look back period of 105 days, such that an episode of MDD would only be considered to be a recurrence if the patient has not had an MDD-related encounter within the past 105 days, thus eliminating routine visits for an ongoing case of MDD from the measure. The 105-day look-back period is an operational provision and not a clinical recommendation, or definition of relapse, remission, or recurrence.

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).
Patients with a suicide risk assessment completed during the visit in which a new diagnosis or recurrent episode was identified

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).
Time Period for Data Collection: At every visit where a new diagnosis or recurrent episode of Major Depressive Disorder is identified [initial evaluation during the episode]

Definition:
Suicide risk assessment - Must include questions about the following:
1) Suicidal ideation
2) Patient´s intent of initiating a suicide attempt
AND, if either is present,
3) Patient plans for a suicide attempt
4) Whether the patient has means for completing suicide

GUIDANCE:
Use of a standardized tool or instrument to assess suicide risk will meet numerator performance. Standardized tools can be mapped
to the concept “Intervention, Performed: Suicide Risk Assessment” included in the numerator logic in the attached HQMF in field S.2a.

HQMF eCQM developed and is attached to this submission in fields S.2a and S.2b.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 18 years and older with a diagnosis of major depressive disorder (MDD)

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).
Time Period for Data Collection: 12 consecutive months

Guidance:
This measure is an episode-of-care measure and should be reported for each instance of a new or recurrent episode of major depressive disorder (MDD); every new or recurrent episode will count separately in the Initial Population.

It is expected that a suicide risk assessment will be completed at the visit during which a new diagnosis is made or at the visit during which a recurrent episode is first identified (ie, at the initial evaluation). For the purposes of this measure, an episode of

MDD would be considered to be recurrent if a patient has not had an MDD-related encounter in the past 105 days. If there is a gap of 105 or more days between visits for MDD, that would imply a recurrent episode. The 105-day look-back period is an operational provision and not a clinical recommendation, or definition of relapse, remission, or recurrence.

The measure description outlined in the header for this measure states, ´patients aged 18 years and older´ while the logic statement states, ´>= 17 year(s) at: "Measurement Period"´. The logic statement, as written, captures patients who turn 18 years old during the measurement period so that these patients are included in the measure. To ensure all patients with major depressive disorder (MDD) are assessed for suicide risk, there are two clinical quality measures addressing suicide risk assessment; CMS 177 covers children and adolescents aged 6 through 17, and CMS 161 covers the adult population aged 18 years and older.

HQMF eCQM developed and is attached to this submission in fields S.2a and S.2b.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
None

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.)*
Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, we encourage the results of this measure to be stratified by race, ethnicity, administrative sex, and payer and have included these variables as recommended data elements to be collected.

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 performance rates:
1. Find the patients who meet the initial population (ie, the general group of patients that a set of performance measures is designed to address).
2. From the patients within the initial population criteria, find the patients who qualify for the denominator (ie, the specific group of patients for inclusion in a specific performance measure based on defined criteria). Note: in some cases the initial population and denominator are identical.
3. From the patients within the denominator, find the patients who meet the numerator criteria (ie, the group of patients in the denominator for whom a process or outcome of care occurs). Validate that the number of patients in the numerator is less than or equal to the number of patients in the denominator

If the patient does not meet the numerator, this case represents a quality failure.

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 is not based on 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 is not based on 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.
Not Applicable

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, Clinician : Individual

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Emergency Department and Services, Other, Outpatient Services
If other: Behavioral Health Day Treatment

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. This measure is not a composite.

2. Validity – See attached Measure Testing Submission Form
Testing_Attachment_MDD_7.1_Final_Intent2Submit.docx,0104_nqf_testing_attachment_7.1_Final-636591402180556115.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0107
Measure Title: Adult Major Depressive Disorder (MDD): Suicide Risk Assessment
Date of Submission: 4/2/2018
Type of Measure:

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

Instructions

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

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

Confidential CMS PQRI 2010 Performance Information by Measure. Jan 2010-Feb 2011 TAP file.

The data source is 2015 EHR data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

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

PQRI/PQRS 2010 data.

The data are for the time period January 2015 through December 2015 and cover the entire United States.

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice

☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

The total number of physicians reporting on this measure, via the EHR reporting option, in 2015, is 380. Of those, 271 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 25,507 quality events. For this measure, 71 percent of physicians are included in the analysis, and the average number of quality reporting events is 94 for the remaining 25,507 events. The range of quality reporting events for 271 physicians included is from 10 to 1,431. The average number of quality reporting events for the remaining 29 percent of physicians that aren't included is 4.

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*).

CMS Physician Quality Reporting Initiative:
Clinical Condition and Measure: #107
27, 665 patients were reported on for the 2010 program, the most recent year for which data are available.

In 2010 the following was reported for this measure:
Eligible Professionals: 108, 484
Professionals Reporting >=1 Valid QDC: 661
% Professionals Reporting >=1 Valid QDC: 0.60%
# Professionals Satisfactorily Reporting: 307
% Professionals Satisfactorily Reporting: 46.40%
Average Reporting Rate per Eligible Professional: 63.30%

There were 25,507 patients included in this reliability testing and analysis. These were the patients that were associated with physicians who had 10 or more patients eligible for this measure.

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.

EHR Measure Validity

- The data sample came from 3 sites representing various types, locations and sizes
  - Site A: A non-profit community mental health center serving over 180 MDD patients per month. The site employs 30 therapists, 5 psychiatrists, 4 nurse practitioners and 1 physician assistant who treat a patient population that is 75% adults and 25% children. The site uses an EHR and all data were extracted electronically. Data was collected from patients seen from 06/1/2012 to 10/31/12.
  - Site B: A solo-private practice in an urban setting serving 5 MDD patients per month. The site uses an EHR and all data were extracted electronically. Data was collected from patients seen from 02/14/2011 to 10/30/12.

    - Site C: A large organization with multiple practice sites in urban and rural settings. The site employs 4,065 physicians serving 445 MDD patients per month. The site uses an EHR in the ambulatory care setting and all data were extracted electronically. Data was collected from patients seen from 07/1/2011 to 6/30/12.
- The sample consisted of 40 charts per site for a total of 120 patients
- Data abstraction was performed between 10/11/2012 and 12/6/2012
- The measure performance was calculated from data collected using two different methods of collection:
    - Automated EHR report
    - Visual inspection of the medical record by professional data abstractors to capture the data elements to manually construct the performance

The same data samples were used for reliability testing and exceptions analysis.

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.

Patient-level socio-demographic (SDS) variables were not captured as part of the testing.

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 analysis included:

- Performance Rate
- Percent agreement for the measure
- Kappa statistic to adjust for chance agreement

Reliability of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in physician performance. Reliability at the level of the specific physician is given by:

Reliability = Variance (physician-to-physician) / [Variance (physician-to-physician) + Variance (physician-specific-error]

Reliability is the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the error variance specific to a physician. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in physician performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the physician performance score is a binomial random variable conditional on the physician's true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates.

Reliability is estimated at two different points, the first is reliability averaged over all the eligible quality reporting events, per provider. The second, includes only those providers that meet the minimum number of quality reporting events for the measure. Each provider must have at least 10 eligible reporting events to be included in this calculation.

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*).

The reliability rates (kappa statistic) between the automated report from the EHR and the manual review of the patient medical records

Reliability: N, % Agreement, Kappa Statistic (95% Confidence Interval)

Overall Reliability: 117, 66.67%, 0.3655 (0.2029, 0.5281)

Denominator Reliability: 120, 99.17%, 0.7959 (0.3976, 1.00)

Numerator Reliability: 117, 66.67%, 0.3655 (0.2029, 0.5281)

The overall average reliability is 0.94. The reliability above the minimum level of quality reporting events is 0.97.

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

The greatest challenge for this measure was that little to none of the patient care performed was documented in a structured, searchable field. More specifically, most patients were found to meet the numerator upon manual review of the patient record because suicide risk assessment was most consistently documented in free text notes by providers. System design improvement efforts could allow for higher reliability for these measures.

This measure has very high reliability when evaluated above the minimum level of quality reporting events, and very high overall reliability.

2b1. VALIDITY TESTING

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

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

☐ **Performance measure score**

☒ **Empirical validity testing**

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

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

Data from a performance report for the measure automatically-generated from the EHR (designed to collect the necessary data elements to identify eligible cases and calculate the performance score) were compared to data elements found and scores calculated manually on visual inspection of the medical record by trained abstractors. 120 patient records were reviewed for this measure.
Data analysis included:

- Performance Rate
- Percent agreement for the measure
- Kappa statistic to adjust for chance agreement

To satisfy NQF's ICD-10 Conversion Requirements, we are providing the information below:

- NQF ICD-10-CM Requirement 1: Statement of intent related to ICD-10 CM
  Goal was to convert this measure to a new code set, fully consistent with the original intent of the measure.

- NQF ICD-10-CM Requirement 2: Coding Table
  See attachment in S.2b

- NQF ICD-10-CM Requirement 3: Description of the process used to identify ICD-10 codes
  The PCPI uses the General Equivalence Mappings (GEMs) as a first step in the identification of ICD-10 codes. We then review the ICD-10 codes to confirm their inclusion in the measure is consistent with the measure intent, making additions or deletions as needed. We have an RHIA-credentialed professional on our staff who review all ICD-10 coding. For measures included in CMS' Quality Payment Program (QPP), the ICD-10 codes have also been reviewed and vetted by the CMS contractor. Comments received from stakeholders related to ICD-10 coding are first reviewed internally. Depending on the nature of the comment received, we also engage clinical experts to advise us as to whether a change to the specifications is warranted.

We conducted a correlation analysis with one other process measure to evaluate empirical validity. Depression Utilization of the PHQ-9 Tool (PQRS #371) was chosen as a suitable candidate for correlation analysis due to the similarities in patient population and domain. We hypothesize that there exists a positive association between patients with major depressive disorder that receive a suicide risk assessment and those that have had a PHQ-9 tool administered.

Datasets were reviewed to identify shared providers based on NPI and TIN identifiers. Correlation analysis was then performed to evaluate the association between performance scores of these shared providers.

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

EHR Measure Validity

The **performance rate was 22.22%** for this measure.

Percent agreement between the manual review and automated report was 99.17% for the denominator of this measure. There was 1 mismatch for the denominator due to an instance when the automated report showed the patient was eligible for the denominator when, in fact, the patient was ineligible during manual review.

Major Depressive Disorder: Suicide Risk Assessment was positively correlated with the Depression Utilization of the PHQ-9 Tool (PQRS #371). Due to the small sample size, the correlation did not reach statistical significance:

PQRS #371
Coefficient of correlation = 0.39
P-value = 0.45

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?*)

Major Depressive Disorder: Suicide Risk Assessment has a moderate and positive correlation with another evidence-based process of care. The moderate correlation demonstrates the criterion validity of the measure.

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*)

N/A.

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*)

N/A.

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*)

N/A.

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.

N/A.

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

N/A.

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.

N/A.

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:

- ☐ **Published literature**
- ☐ **Internal data analysis**
- ☐ **Other (please describe)**

N/A.

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

N/A.

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

N/A.

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

N/A.

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

N/A.

2b3.9. Results of Risk Stratification Analysis:

N/A.

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*)

N/A.

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*)

N/A.

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)*

CMS Physician Quality Reporting Initiative:
The inter-quartile range (IQR) was calculated, which provides a measure of the dispersion of performance.

Measures of central tendency, variability, and dispersion were calculated.

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*)

2015 Physician Quality Reporting System (PQRS) Experience Report

2015 is the most recent year for which PQRS Experience Report measure data are available. The average performance rates on Adult Major Depressive Disorder (MDD): Suicide Risk Assessment over the last several years are as follows:

- 2012: 77.0%
- 2013: 76.7%
- 2014: 86.0%
- 2015: 71.3%

2015 Reporting Experience, Including Trends (2007-2016), Physician Quality Reporting System. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/AnalysisAndPayment.html

It is important to note that PQRS has been and remains a voluntary reporting program. In the early years of the PQRS program,

participants received an incentive for satisfactorily reporting. As a result, performance rates may not be nationally representative. Beginning in 2015, the program imposed payment penalties for non-participants based on 2013 performance

Based on the sample of 271 included physicians, the mean performance rate is 0.68 the median performance rate is 0.86 and the mode is 1.0. The standard deviation is 0.34. The range of the performance rate is 0.99 with a minimum rate of 0.01 and a maximum rate of 1.0. The interquartile range is 0.56 (0.96–0.40).

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 range of performance from 0.01 to 1 suggests there's clinically meaningful variation across physicians' performance.

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*)

This test was not performed for this measure.

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*)

This test was not performed for this measure.

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*)

This test was not performed for this measure.

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*).

Data are not available to complete this testing.

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*)

Data are not available to complete this testing.

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*)

Data are not available to complete this testing.

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 by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition
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).

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 greatest challenge for this measure was that little to none of the patient care performed was documented in a structured, searchable field. More specifically, most patients were found to meet the numerator upon manual review of the patient record because suicide risk assessment was most consistently documented in free text notes by providers. System design improvement efforts could allow for higher reliability for these measures.

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)*.
The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain.

Commercial uses of the Measures require a license agreement between the user and the PCPI® Foundation (PCPI®) or the American Medical Association (AMA). Neither the American Medical Association (AMA), nor the AMA-convened Physician Consortium for Performance Improvement® (AMA-PCPI), now known as the PCPI®, nor their members shall be responsible for any use of the Measures.

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 MIPS https://qpp.cms.gov/mips/quality-measures

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

Merit-based Incentive Payment System (MIPS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)
Prior to 2016, this measure was used for Eligible Providers (EPs) in the Physician Quality Reporting System (PQRS). As of 2017, PQRS has been replaced by the Merit-based Incentive Payment System (MIPS). MIPS is a national performance-based payment program that uses performance scores across several categories to determine payment rates for EPs. MIPS takes a comprehensive approach to payment by basing consideration of quality on a set of evidence-based measures that were primarily developed by clinicians, thus encouraging improvement in clinical practice and supporting advances in technology that allow for easy exchange of information.

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?*)
According to the CY 2018 Quality Payment Program final rule, CMS intends to "make all measures under MIPS quality performance category available for public reporting on Physician Compare in the transition year of the Quality Payment Program, as technically feasible." These measures include those reported via all available submission methods for MIPS-eligible clinicians and groups. Because this measure has been in use for at least one year and meets the minimum sample size requirement for reliability, this measure meets criteria for public reporting. 2018 data will be available for public reporting on Physician Compare in late 2019.

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.*)

Because this measure has been in use for at least one year and meets the minimum sample size requirement for reliability, this measure meets criteria for public reporting. 2018 data will be available for public reporting on Physician Compare in late 2019.

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.
The PCPI measure development process is a rigorous, evidence-based process that has been refined and standardized over the past fifteen years, since the PCPI's inception. Throughout its tenure, several key principles have guided the development of performance measures by the PCPI, including the following which underscore the role those being measured have played in the development process and later through implementation feedback :

Collaborative Approach to Measure Development
PCPI measures have been developed through cross-specialty, multi-disciplinary expert work groups. Representatives of all relevant
disciplines of medicine and other health care professionals are invited to participate as equal contributors to the measure development process. In addition, the PCPI strives to include on its work groups individuals representing the perspectives of patients, consumers, private health plans, and employers. Liaisons from key measure development organizations, including The Joint Commission and NCQA participate in the PCPI's measure development process to ensure harmonization of measures; measure methodologists, coding and informatics experts also are considered important members of the work group. This broad-based approach to measure development maximizes measure buy-in from stakeholders and minimizes bias toward any individual specialty or stakeholder group. As noted in Ad.1 below, 22 individuals from a diverse group of specialties including psychiatry, family medicine, nursing, occupational therapy, social work, internal medicine, and psychology contributed to the development of this measure.

Conduct Public Comment Period
Input from multiple stakeholders is integral to the measure development process. In particular, feedback is critical from those clinicians who will implement these measures.. To that end, all measures are released for a 30-day public and PCPI member comment period. All comments are reviewed by the work group to determine whether measure modifications are needed based on comments received.

Feedback Mechanism
The PCPI has a dedicated process set up to receive comments and questions from implementers. As comments and questions are received, they are shared with appropriate staff for follow up. If comments or questions require expert input, these are shared with the PCPI's expert work groups to determine if measure modifications may be warranted. Additionally, for PCPI measures included in federal reporting programs, there is a system that has been set up to elicit timely feedback and responses from PCPI staff in consultation with work group members, as appropriate.

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.**
See description in 4a1.1 above.

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.
In addition to the feedback obtained from cross-specialty, multi-disciplinary work groups during the measure development process, the PCPI obtains feedback via a public comment period and an email-based process set up to receive measure inquiries from implementers. The public comment period feedback is provided via an online survey tool and, as mentioned, implementer feedback is provided via email.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
The most salient theme during the public comment period was that the original measure required a more complex assessment of suicide risk that could deter non-mental health providers treating depression from reporting on this important measure. It was suggested to reduce the complexity of the assessment to include the most essential elements in the assessment of suicide risk.

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

As stated above, we received feedback from measure implementers that it was difficult to capture only the new and new recurrent episodes of MDD.

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.**
Based on feedback and recommendations from those being measured, we reduced the original number of suicide risk assessment components to the four most essential in the suicide risk assessment. This change was intended to reduce the complexity of the measure and make it easier to for all providers who treat patients with depression to report on.

Based on feedback from measure implementers, we modified the measure logic to include a lookback period for a prior diagnosis of new or recurrent MDD to ensure that routine visits for an ongoing case of MDD (which do not require a suicide risk assessment) were not included in the measure.

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.
While the PCPI creates measures with an ultimate goal of improving the quality of care, measurement is a mechanism to drive improvement but does not equate with improvement. Measurement can help identify opportunities for improvement with actual improvement requiring making changes to health care processes and structure. In order to promote improvement, quality measurement systems need to provide feedback to front-line clinical staff in as close to real time as possible and at the point of care whenever possible. (1)

1. Conway PH, Mostashari F, Clancy C. The future of quality measurement for improvement and accountability. JAMA. 2013 Jun 5;309(21):2215-6.

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.**
We are not aware of any unintended consequences related to this measurement.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
We are not aware of any unexpected benefits from implementation 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)**

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

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.**
The guidelines used as evidence in the NQF 1365: Child and Adolescent Major Depressive Disorder (MDD) Suicide Risk Assessment explicitly recommend suicide assessment at every visit for MDD whereas the guidelines used for evidence in this measure do not emphasize this level of assessment frequency.

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.)
Both of these measures (0104 and 1365) were developed by PCPI and updated and harmonized with each other on an annual basis. They are not competing because they are used in different patient populations and have different frequencies of suicide assessment based on their respective evidence.

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):** PCPI
Co.2 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@thepcpi.org, 312-224-6071-
Co.3 **Measure Developer if different from Measure Steward:** PCPI
Co.4 **Point of Contact:** Courtney, Hurt, courtney.hurt@thepcpi.org, 312-224-6069-

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.

PCPI measures are developed through cross-specialty, multi-disciplinary technical expert panels (TEPs). Representatives of all relevant disciplines of medicine and other health care professionals are invited to participate. In addition, the PCPI strives to include on its TEPs individuals representing the perspectives of patients, consumers, private health plans, and employers. Measure methodologists, and coding and informatics experts also are considered important members of the TEP. All TEP members participate as equal contributors to the measure development process. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. TEPs were convened in 2001 and 2010 to develop, refine and maintain a set of measures addressing mental health including measure #0104. More recently, in 2016, the PCPI reconvened the Mental Health TEP which included the following individuals.

John Absher, MD (neurology)
Alan Axelson, MD (psychiatry)
Andrea Bostrom, PhD, PMHCNS-BC (nursing, psychiatric nursing)
Mirean Coleman, MSW, LICSW, CT (social work)
Mary Dobbins, MD (psychiatry)
Mary Ann Forciea, MD (internal/geriatric medicine)
Elizabeth M. Galik, PhD, CRNP (nursing)
Jerry Halverson, MD (psychiatry, methodology)
Richard Hellman, MD, FACP, FACE (endocrinology, methodology)
Renee Kinder, MS, CCC-SLP (rehabilitation, gerontology)
Helen H. Kyomen, MD, MS (geriatric and adult psychiatry)
Katie Maslow, MSW (patient advocacy representative)
John S. McIntyre, MD, DFAPA, FACPsych (psychiatry, methodology)
Karen Pierce, MD (psychiatry)
Joseph W. Shega, MD (geriatric medicine, hospice and palliative medicine)
Eric G. Tangalos, MD, FACP, AGSF, CMD (internal/geriatric medicine)
Roberta Waite, EdD, APRN, CNS-BC (psychiatric nursing, methodology)

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2006
Ad.3 **Month and Year of most recent revision:** 05, 2017
Ad.4 **What is your frequency for review/update of this measure?** Supporting guidelines, specifications, and coding for this measure are reviewed annually
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2018

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** Coding/Specifications updates occur annually. The PCPI has a formal measurement review process that stipulates regular (usually on a three-year cycle, when feasible) review of the measures. The process can also be activated if there is a major change in scientific evidence, results from testing or other issues are noted that materially affect the integrity of the measure.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0105
Measure Title: Antidepressant Medication Management (AMM)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of members 18 years of age and older who were treated antidepressant medication, had a diagnosis of major depression, and who remained on an antidepressant medication treatment. Two rates are reported.

a) Effective Acute Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 84 days (12 weeks).
b) Effective Continuation Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 180 days (6 months).

a) Effective Acute Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 84 days (12 weeks).
b) Effective Continuation Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 180 days (6 months).
Developer Rationale: Clinical guidelines for depression emphasize the importance of effective clinical management in increasing patients' medication compliance, monitoring treatment effectiveness, and identifying and managing side effects. If pharmacological treatment is initiated, appropriate dosing and continuation of therapy through the acute and continuation phases decrease recurrence of depression. Thus, evaluation of duration of pharmacological treatment serves as an important indicator in promoting patient compliance with the establishment and maintenance of an effective medication regimen.

Numerator Statement: Adults 18 years of age and older who were newly treated with antidepressant medication, had a diagnosis of major depression, and who remained on an antidepressant medication treatment.
Denominator Statement: Patients 18 years of age and older with a diagnosis of major depression and were newly treated with antidepressant medication.
Denominator Exclusions: Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients who did not have a diagnosis of major depression in an inpatient, outpatient, ED, telehealth, intensive outpatient or partial hospitalization setting during the 121-day period from 60 days prior to the IPSD, through the IPSD and the 60 days after the IPSD.

Exclude patients who filled a prescription for an antidepressant 105 days prior to the IPSD.

Measure Type: Process
Data Source: Claims, Electronic Health Data
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** February 28, 2014

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence

Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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 that links continuation of antidepressant medications to less episodes of major depression and lower morbidity.
- The developer includes updates to the evidence in this submission including clinical practice guidelines and systematic reviews:
  - Practice guideline for the treatment of patients with major depressive disorder. American Psychiatric Association. Oct 2010/Oct 2015. Within the guideline, 4 randomized double-blind clinical trials, 1 clinical trial, and 1 qualitative review were referenced. **All recommendations received a [I] grade (recommended with substantial clinical confidence).**
  - VA/DoD Management of Major Depressive Disorder in Adults in the Primary Care Setting. Department of Veterans Affairs, and Health Affairs, Department of Defense. April 2016. The guideline cited 2 Random Control Trials (RCT), 2 systematic reviews, and 1 clinical study. **The recommendations were graded "strong" or "weak". 2009 Clinical Practice Guideline for Management of Major Depressive Disorder based on evidence reviewed through 2007 was graded C and B.**
  - Institute for Clinical Systems Improvement: Recommendations for the Diagnosis and Treatment of Major Depression in Adults in Primary Care. Trangle, M., et al. March 2016. **This guideline includes 3 studies showing "high level" evidence, 1 systematic review, and 5 studies showing "low level" evidence.**
  - Antidepressant Drug effects and Depression Severity: A Patient Level Meta-Analysis. Fournier, J., et al. January, 2010. **This review included five RCTs. The evidence was not graded.**
  - Antidepressants for treatment of depression in primary care: a systematic review and meta-analysis. Arroll, B., et al. December 2016. This review included 17 RCTs. **The evidence was not graded.**
- The developer notes several studies in addition to the guidelines above: 2 RCTs; 1 qualitative reviews; 2 prospective studies; 2 survey studies; 1 case study; and 1 fact sheet.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provides updates to previous guidelines and systematic reviews as well as one new systematic review listed above.

Exception to evidence:

- N/A

Questions for the Committee:

- *The evidence provided by the developer has been updated and is directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

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) → High (Box 5a) → High

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 performance data at the health plan level (HEDIS) demonstrating variation in performance and opportunity for improvement.
- The summarized health plan data is stratified by year and product line (i.e. commercial, Medicare, Medicaid).

Disparities

- The developer cites disparities data from several studies showing disparities in treatment for major depressive episodes:
  - The percentage of adults with a major depressive episode in 2008, who received treatment for it, was significantly lower for blacks than for whites (58.9 vs. 71.1 percent) and for Hispanics than non-Hispanic whites (51.8 vs. 73.3 percent). (AHRQ, 2009)
  - A study examining antidepressant treatment patterns found that, compared to younger adults, older adults tended to be more likely to discontinue antidepressant treatment (Sanglier et al., 2011).
  - A study that examined the treatment disparities for respondents with major depressive disorders showed that blacks and Hispanics were less likely to use antidepressants than whites. (Fleming et al., 2003).
  - Compared to whites, blacks and Hispanics in primary care were less likely to be prescribed antidepressants for their depression. Whites also received more antidepressant prescriptions after a visit to psychiatrists when compared to blacks (Lagomasino et al., 2011).

Questions for the Committee:

- *Does the Committee agree that the updated performance data demonstrate 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 is evidence to support use and rationale for this measure. This seems to be an easy means to capture whether there is adherence to newly prescribed antidepressants. Measure reflects adherence at 12 weeks and 6 months after prescription initiated.
**A number of updates to the original evidence submitted when this measure was submitted were presented, e.g., APA guidelines were updated, a number of new guidelines created as well as additional RCTs and studies. Evidence, including new evidence, was rated high. Medication management for individuals using antidepressant medications is a critical component of comprehensive treatment of depression is a clinical.
**Clearly all the guidelines and significant literature support using antidepressants reliably. This measure is correlated with improved outcomes (remission and response) but not nearly as closely or directly as the other NQF endorsed measures which track PHQ-9 use, response rates and remission rates. It also excludes all patients that receive psychotherapy only. At what point do we consider de-commissioning measures like this that clutter the landscape and are less direct and core to desired outcomes?
**Process measure.

1b. Performance Gap
Comments:
**There is considerable variability in performance suggesting that most health plans are not strongly advocating and educating on necessity of adherence to treatment. Disparities in care highlight fewer prescriptions for Blacks and Latinos and less adherence when they are prescribed.
**The performance gap is disturbing. Improvement on this measure which is used for accountability by more than 7 quality programs including the Medicaid Adult Core Set, MIPS and HEDIS, is 1% over the last 3 years; highest for the Medicare population but with a shocking 17% lower performance gap for Medicaid patients suggesting the need for performance incentives for that population. There are significant disparities with blacks and Hispanics were less likely to receive medication management for antidepressants and less likely to use medications.
**Gap clearly exists but I care more about outcomes gap and less about whether a patient is taking a pill. I can take a pill regularly but remain deeply depressed.
**A performance gap exists. Disparities are evidence with Blacks and Hispanics receiving less treatment for a major depressive episode.

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.

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)?*
- *The 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 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**

Committee pre-evaluation comments

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

2a1. Reliability – Specifications

Comments:

**This measure should have no challenges being consistently implemented. Medications are clearly identified and calculation logic is clear.
**No concerns overall but a question: the numerator for this measure includes the following: "patients who remained on medication." What does this mean? For how long --the entire period used for the measure? something else? The term is not defined anywhere.
**It's reliable.
**Reliability is high.

2a2. Reliability – Testing

Comments:

**My only concern is whether it is not broad enough and should include all who are prescribed an antidepressant. Since it isn't only mental health professionals making MDD diagnosis, chances are diagnosis will not be 100% accurate and will miss people or include people without true MDD. For these reasons, I think 12 week/6 month adherence to antidepressant should be reflected as the measure indicator absent MDD diagnosis.
**No concerns.
**No.

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

Comments:

**I am a bit confused why all prescriptions for an antidepressant, regardless of MDD diagnosis aren't to be monitored similarly? In the event the provider sees need for patient to be on an antidepressant (should sxs be approaching MDD or some other mood disorder), adherence would be equally as important.
**No concerns
**It's valid
**Moderate validity.

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

Comments:

**No risk adjustment or risk stratification
**Process measure
**N/A

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 defined fields in a combination of electronic sources.
- No fees or licensure requirements are required.
- The developer notes that the measure has precise specifications but data methods and calculation methods may vary. Therefore, NCQA conducts an independent audit in order to verify that HEDIS specifications are met.

Questions for the Committee:

o *Does the Committee have any concerns in regards to the feasibility of the measure?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**Data elements can feasibly be generated. Only piece that will be less reliable is MDD diagnosis as this is clinician generated and doesn't require standard tool.
**No special concerns All data elements are in electronic records in one place or another.
**It's feasible.
**Feasibility is good--data elements are in electronic records.

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**

Accountability program details

The developer reports that the measure is used in the following programs:

- Medicaid Adult Core Set;
- Merit Based Incentive Payment System (MIPS) Quality Payment Program (QPP);
- Health Insurance Exchange Quality Rating System (QRS);
- State of Health Care Annual Report;
- Health Plan Rating/Report Cards;
- Health Plan Accreditation

- Quality Compass

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

- The developer provides performance results and data annually in their Quality Compass tool and presents data at various conferences and webinars. The developer also provides regular technical assistance through its Policy Clarification Support System.
- The developer uses several methods to obtain input from users during its "reevaluation process", including, vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer noted that the health plans have not reported significant implementation barriers.

Additional Feedback:

- N/A

***Questions for the Committee**:*

o *Are the methods the developer used to vet the measure sufficient?*

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 notes a slight improvement (approximately an one percentage point increase) across health plans over the past three years. The Medicare population showed the highest performance for both the acute and continuation indicators. The Medicaid population shows the largest gap in performance, averaging approximately 17 percentage points lower than Medicare.

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.

***Questions for the Committee**:*

o *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:

**Used in 7 major accountability data sets. Feedback is received in multiple ways.
**No problems here
**Publicly reported data.

4b1. Usability – Improvement

Comments:

**From reviewing this submission, I can't tell what happens if patient's antidepressant is switched from one to another or if that rolls into a new 12 week/6 month course of treatment. Should that be the case, then it would appear that adherence is reduced and I wouldn't want physicians to stick with compliance in order to meet this measure.
**No harms have been noted and there should be substantial benefits. However improvement has been 1% over the past 3 years in HEDIS and it is clear from the data that the Medicaid population needs some special attention and incentives need to be provided to assure performance improvement in the use of medication and its management by the Medicaid population.
**No problems here.
**There has been slight improvement---one percent increase across health plans.

Criterion 5: Related and Competing Measures

Related or competing measures

- #1880 – Adherence to Mood Stabilizers for People with Bipolar I Disorder.

Harmonization

- Harmonization plan not submitted.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 0105
Measure Title: Antidepressant Medication Management

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document

   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.

   This measure is specified at the health plan level of analysis. Claims data were used for testing.

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

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐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, skip Questions #3-8, then go to Question #9)

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

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

The dataset included 2016 Healthcare Effectiveness Data and Information Set (HEDIS) data. The developer calculated measure score reliability using 2016 HEDIS data that included 401 Medicare health plans, 226 Medicaid health plans, and 403 commercial health plans.

☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

REFERENCE: Testing attachment, section 2a2.2
***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 #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

The developer used a beta-binomial model to calculate the signal to noise ratio.

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

REFERENCE: Testing attachment, section 2a2.2
***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 #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

Results of the reliability testing:

Beta-binomial statistic for each measure rate:

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

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

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #8)

   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?

   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

   ☐Insufficient (go to Question #9)

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

   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    ***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*
    ☒Yes (go to Question #12)
    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

12. Analysis of potential threats to validity: Any concerns with measure exclusions?
    REFERENCE: Testing attachment, section 2b2.
    ***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 #13)
    ☐No (go to Question #13)
    ☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
    REFERENCE: Testing attachment, section 2b3.

    13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

    13b. Are social risk factors included in risk model? ☐Yes ☐No

    13c. Any concerns regarding the risk-adjustment approach?
    ***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*
    ☐Yes (please explain below then go to Question #14)
    ☐No (go to Question #14)
    ☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
    REFERENCE: Testing attachment, section 2b4.
    ☐Yes (please explain below then go to Question #15)
    ☒No (go to Question #15)

The developer calculated an inter-quartile range (IQR) for the acute and continuation phases for all product lines (i.e., commercial, Medicare and Medicaid plans). The difference between the 25th and 75th percentile for each phase of the product lines is <0.001, which shows the plans are significantly different from each other.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)
The developer conducted both construct and face validity.

Construct validity was tested using the Pearson correlation coefficient to assess whether the Antidepressant Medication Management measure correlated with Statin Therapy for Patients with Diabetes measure in Medicare, commercial, and Medicaid plans. The hypothesis was that organizations that perform well on Antidepressant Medication Management measure should perform well on the Statin Therapy measure given that the measures are about health plan's success in improving adherence to medication treatment for chronic conditions.

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***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 #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

Testing results show that the Antidepressant Medication Management measure is positively correlated with the Statin Therapy for Patients With Diabetes measure across all three plans: Medicaid (correlation coefficient for acute phase is 0.50 and continuation phase is 0.49); Commercial (correlation coefficient for the acute phase is 0.69 and continuation phase is 0.69); and Medicare plans (correlation coefficient for the acute phase is 0.56 and continuation phase is 0.60).

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

REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)
☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

***REFERENCE:** Testing attachment, section 2b1.*
***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 #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.

***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 #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

The developer conducted both empirical and face validity testing. The results from the multi-stakeholder advisory panel indicated the measure will accurately differentiate quality across providers.

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0105
Measure Title: Antidepressant Medication Management
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: Click here to enter a date

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: Continuation of antidepressant medications for people newly treated with medications

☐ 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.
Measure continuation of antidepressant medication >> Identify people diagnosed with major depression who were recently prescribed an antidepressant medication >> Assess adherence to medication within the acute and continuation phases of treatment >> Identify people who are not continuing their pharmacotherapy >> Improve rates of relapse by focusing on improving adherence to antipsychotics for people who begin treatment >> Less episodes of major depression and lower morbidity

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

****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

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: Clinical Practice Guideline 1

Source of Clinical Practice Guideline: o **Title**	o Practice guideline for the treatment of patients with major depressive disorder, third edition o American Psychiatric Association (APA)

o **Author** o **Date** o **Citation, including page number** o **URL**	o October 2010. (The American Psychiatric Association reaffirmed the currency of the guideline in October 2015.) o American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder, third edition. Arlington (VA): American Psychiatric Association; 2010 Oct. p. 152 o https://www.guideline.gov/summaries/summary/24158/Practice-guideline-for-the-treatment-of-patients-with-major-depressive-disorder-third-edition
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.	•~~"Successful treatment of patients with major depressive disorder is promoted by a thorough assessment of the patient and close adherence to treatment plans. Treatment consists of an acute phase, during which remission is induced; a continuations phase, during which remission is preserved; and a maintenance phase, during which the susceptible patients is protected against the recurrence of a subsequent major depressive episode."~~ •"An antidepressant medication is recommended as an initial treatment choice for patients with mild to moderate major depressive disorder [I: Recommended with substantial clinical confidence] and definitely should be provided for those with severe major depressive disorder unless electroconvulsive therapy (ECT) is planned [I: Recommended with substantial clinical confidence]." •Patients should be given a realistic notion of what can be expected during the different phases of treatment, including the likely time course of symptom response and the importance of adherence for successful treatment and prophylaxis [I]. •During the acute phase of treatment, patients should be carefully and systematically monitored on a regular basis to assess their response to pharmacotherapy, identify the emergence of side effects (e.g., gastrointestinal symptoms, sedation, insomnia, activation, changes in weight, and cardiovascular, neurological, anticholinergic, or sexual side effects), and assess patient safety [I]. •"During the continuation phase of treatment, the patient should be carefully monitored for signs of possible relapse [I: Recommended with substantial clinical confidence]. Systematic assessment of symptoms, side effects, adherence, and functional status is essential [I: Recommended with substantial clinical confidence], and may be facilitated through the use of clinician- and/or patient-administered rating scales [II: Recommended with moderate clinical confidence]. To reduce the risk of relapse, patients who have been treated successfully with antidepressant medications in the acute phase should continue treatment with these agents for 4–9 months [I: Recommended with substantial clinical confidence]."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	"The type of evidence supporting the recommendations is not specifically stated. In order for the reader to appreciate the evidence base behind the guideline recommendations and the weight that should be given to each recommendation, the summary of treatment recommendations is keyed according to the level of confidence with which each recommendation is made (see "Major Recommendations" field). Each rating of clinical confidence considers the strength of the available evidence. When evidence from randomized controlled trials and meta-analyses is limited, the level of

	confidence may also incorporate other clinical trials and case reports as well as clinical consensus with regard to a particular clinical decision." All recommendations above received a [I] grade (Recommended with substantial clinical confidence) ~~The pharmacotherapy recommendations received a [I] grade (Recommended with substantial clinical confidence)~~
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	Pharmacotherapy recommendations received a [I] grade (Recommended with substantial clinical confidence)
Provide all other grades and definitions from the recommendation grading system	APA RATING SCHEME FOR THE STRENGTH OF THE RECOMMENDATION Each recommendation is identified as falling into one of three categories of endorsement, indicated by a bracketed Roman numeral following the statement. The three categories represent varying levels of clinical confidence: [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?	Quantity: Within the APA guideline, recommendations specific to pharmacotherapy adherence reference 4 randomized double-blind clinical trials, 1 clinical trial, and 1 qualitative review. Here are some examples of the studies referenced by the APA guideline. One randomized double-blind trial (Keller, 1998) looked at 635 outpatients at 12 sites who met criteria for major depression. Another randomized double-blind trail (Keller, 2007) included 1096 outpatients who were offered two different types of antidepressants to examine the effect of mediation on the prevention of recurring depressive episodes. A meta-analysis (Hansen, 2008) was conducted of RCTs, meta-analyses and observational studies published between 1980 and 2007 and found an overall benefit to continuation and maintenance of antidepressant pharmacotherapy. Quality: The APA Guideline recommends with substantial clinical confidence that people with mild to major depression should adhere to appropriate pharmacotherapy (antidepressants).
Estimates of benefit and consistency across studies	The benefit of adherence to antidepressant medications is a reduction in the recurrence rate of new episodes of depression. The guidelines and evidence note that pharmacotherapy is most effective when the physician identifies the most effective medication for each patient. This performance measure focuses on continuation of medication during the acute and continuation phases of treatment. Evidence suggests that physicians can help maximize the efficacy of medication treatment by monitoring the effects of the medication and the dosage. Across included studies, guidelines agree that antidepressants are an effective way to treat people with major depression, if steps are taken to help patients adhere to their medications.

What harms were identified?	The guidelines and evidence note that pharmacotherapy is most effective when the physician identifies the most effective medication for each patient. The harms stem from a lack of adherence to medications.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 2: Clinical Practice Guideline 2

Source of Clinical Practice Guideline: o **Title** o **Author** o **Date** o **Citation, including page number** o **URL**	o Management of Major Depressive Disorder in Adults in the Primary Care Setting o Department of Veterans Affairs, and Health Affairs, Department of Defense o May 2000 o Management of Major Depressive Disorder in Adults in the Primary Care Setting. Washington, DC: VA/DoD Evidence Based Clinical Practice Guideline Working Group, Veterans Health Administration, Department of Veterans Affairs, and Health Affairs, Department of Defense; May 2000. Office of Quality and Performance publication 10Q-CPG/MDD-00. o http://www.oqp.med.va.gov/cpg/MDD/MDD_Base.htm o VA/DoD Clinical Practice Guideline for the Management of Major Depressive Disorder o The Department of Veterans Affairs and the Department of Defense o April 2016 o VA/DoD Clinical Practice Guideline for the Management of Major Depressive Disorder. Washington, DC: VA/DoD Evidence-Based Practice Working Group, Veterans Health Administration, Department of Veterans Affairs, and Health Affairs, Department of Defense; April 2016. Office of Quality and Performance publication o https://www.healthquality.va.gov/guidelines/MH/mdd/VADoDMDDCPGFINAL82916.pdf NOTE: In 2009, the VA and DoD published a Clinical Practice Guideline (CPG) for the Management of Major Depressive Disorder (2009 MDD CPG), which was based on evidence reviewed through 2007. The current document is an update to the 2009 MDD CPG. The CPG states: "The MDD CPG Work Group focused largely on developing new and updated recommendations based on the evidence review conducted for the priority areas addressed by the key questions. In addition to those new and updated recommendations, the CPG Work Group considered the current applicability and relevance of the remaining recommendations that were made in the previous 2009 MDD CPG. While these remaining 2009 recommendations were reviewed by the group, the literature supporting these

	recommendations was not reviewed as part of a systematic literature search. Therefore, the determination of carrying forward or modifying these prior recommendations was based on expert opinion as well as on the evidence review from the previous version of the guideline. In order to be fully transparent, Appendix F [recommendation table] displays all the recommendations from the 2009 MDD CPG and the information regarding how 2009 recommendations were incorporated into the 2016 MDD CPG, including the recommendation category and the 2016 recommendation to which it corresponds, if applicable." We have included both the 2009 and 2016 grades/categories for each recommendation included below.
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.	"When antidepressant pharmacotherapy is used, the following key messages should be given to enhance adherence to medication: [B: A recommendation that clinicians provide (the service) to eligible patients.] •Most people need to be on medication for at least 6 to 12 months after adequate response •It usually takes 2 to 6 weeks before improvements are seen •Continue to take the medication even after feeling better •Do not discontinue taking medications without first discussing with your provider" 1. "As first-line treatment for uncomplicated mild to moderate MDD, we recommend offering one of the following treatments based on patient preference, safety/side effect profile, history of prior response to a specific medication, family history of response to a medication, concurrent medical illnesses, concurrently prescribed medications, cost of medication and provider training/competence: •Evidence-based psychotherapy: • Acceptance and commitment therapy (ACT) • Behavioral therapy/behavioral activation (BT/BA) • Cognitive behavioral therapy (CBT) • Interpersonal therapy (IPT) • Mindfulness-based cognitive therapy (MBCT) • Problem-solving therapy (PST) •Evidence-based pharmacotherapy: • Selective serotonin reuptake inhibitor (except fluvoxamine)(SSRIs) • Serotonin–norepinephrine reuptake inhibitor (SNRIs) • Mirtazapine • Bupropion • The evidence does not support recommending a specific evidence-based psychotherapy or pharmacotherapy over another." [2009 Evidence Grade: A, B. 2016 Grade: Strength: Strong For, Category: Reviewed, New-replaced] 2. We suggest offering a combination of pharmacotherapy and evidence-based psychotherapy for the treatment of patients with MDD during a new episode of care when the MDD is characterized as: • Severe (i.e., PHQ-9 >20)

	• Chronic (duration greater than two years) • Recurrent (with three or more episodes)" [2009 Evidence Grade: A. 2016 Grade: Strength: Weak For, Category: Reviewed, New-replaced] 3. "In patients who have demonstrated partial or no response to initial pharmacotherapy monotherapy (maximized) after a minimum of four to six weeks of treatment, we recommend switching to another monotherapy (medication or psychotherapy) or augmenting with a second medication or psychotherapy." [2009 Evidence Grade: None. 2016 Grade: Strength: Strong For, Category: Reviewed, New-replaced] 4. "After initiation of therapy or a change in treatment, we recommend monitoring patients at least monthly until the patient achieves remission. At minimum, assessments should include a measure of symptoms, adherence to medication and psychotherapy, and emergence of adverse effects." [2009 Evidence Grade: C, B. 2016 Grade Strength: Strong For, Category: Reviewed, Amended]
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	~~Pharmacotherapy continuation receive an [A] grade~~ Note: As explained above, the evidence review for these recommendations is from the 2009 CPG. The evidence was graded using the USPSTF evidence grading system. 1. 2009 Evidence Review Grade: A- a strong recommendation that the clinicians provide the intervention to eligible patients. Good evidence was found that the intervention improves important health outcomes and concludes that benefits substantially outweigh harm.; B- a recommendation that clinicians provide (the service) to eligible patients. At least fair evidence was found that the intervention improves health outcomes and concludes that benefits outweigh harm. 2. 2009 Evidence Review Grade: A- a strong recommendation that the clinicians provide the intervention to eligible patients. Good evidence was found that the intervention improves important health outcomes and concludes that benefits substantially outweigh harm. 3. Evidence not graded. 4. 2009 Evidence Review Grade: B- a recommendation that clinicians provide (the service) to eligible patients. At least fair evidence was found that the intervention improves health outcomes and concludes that benefits outweigh harm.; C- no recommendation for or against the routine provision of the intervention is made. At least fair evidence was found that the intervention can improve health outcomes, but concludes that the balance of benefits and harms is too close to justify a general recommendation.
Provide all other grades and definitions from the evidence grading system	D: Recommendation is made against routinely providing the intervention to asymptomatic patients. At least fair evidence was found that the intervention is ineffective or that harms outweigh benefits.
Grade assigned to the **recommendation**	Note: As explained above, the recommendation grade (including both a "strength" and "category") for these recommendations was updated in the 2016 CPG, while the evidence was reviewed in 2009.

with definition of the grade	1. 2016 Grade: Strength: Strong For, Category: Reviewed, New-replaced. The CPG recommends offering this option for care. Recommendation from previous CPG that has been carried over to the updated CPG that has been changed following review of the evidence. 2. 2016 Grade: Weak For, Category: Reviewed, New-replaced. The CPG suggests offering this option for care. Recommendation from previous CPG that has been carried over to the updated CPG that has been changed following review of the evidence. 3. 2016 Grade: Strength: Strong For, Category: Reviewed, New-replaced. The CPG recommends offering this option for care. Recommendation from previous CPG that has been carried over to the updated CPG that has been changed following review of the evidence. 4. 2016 Grade Strength: Strong For, Category: Reviewed, Amended. The CPG recommends offering this option for care. Recommendation from the previous CPG that has been carried forward to the updated CPG where the evidence has been reviewed and a minor amendment has been made
Provide all other grades and definitions from the recommendation grading system	~~VA/DOD RATING SCHEME FOR THE STRENGTH OF THE RECOMMENDATION~~ ~~A: A strong recommendation that the clinicians provide the intervention to eligible patients.~~ ~~Good evidence was found that the intervention improves important health outcomes and concludes that benefits substantially outweigh harm.~~ ~~B: A recommendation that clinicians provide (the service) to eligible patients.~~ ~~At least fair evidence was found that the intervention improves health outcomes and concludes that benefits outweigh harm.~~ ~~C: No recommendation for or against the routine provision of the intervention is made.~~ ~~At least fair evidence was found that the intervention can improve health outcomes, but concludes that the balance of benefits and harms is too close to justify a general recommendation.~~ ~~D: Recommendation is made against routinely providing the intervention to asymptomatic patients.~~ ~~At least fair evidence was found that the intervention is ineffective or that harms outweigh benefits.~~ ~~I: The conclusion is that the evidence is insufficient to recommend for or against routinely providing the intervention.~~ ~~Evidence that the intervention is effective is lacking, or poor quality, or conflicting, and the balance of benefits and harms cannot be determined.~~ The relative strength of the recommendation is based on a binary scale, "Strong" or "Weak." A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh

	undesirable outcomes. If the Work Group is less confident of the balance between desirable and undesirable outcomes, they present a weak recommendation. Similarly, a recommendation for a therapy or preventive measure indicates that the desirable consequences outweigh the undesirable consequences. A recommendation against a therapy or preventive measure indicates that the undesirable consequences outweigh the desirable consequences. Using these elements, the grade of each recommendation is presented as part of a continuum: • Strong For (or "We recommend offering this option …") • Weak For (or "We suggest offering this option …") • Weak Against (or "We suggest not offering this option …") • Strong Against (or "We recommend against offering this option …") Additional Recommendation Categories and Definitions • Reviewed- New-added: New recommendation following review of the evidence • Reviewed- Not changed: Recommendation from previous CPG that has been carried forward to the updated CPG where the evidence has been reviewed but the recommendation is not changed • Reviewed- Deleted: Recommendation from the previous CPG that has been removed based on review of the evidence • Not reviewed- Not changed: Recommendation from previous CPG that has been carried forward to the updated CPG, but for which the evidence has not been reviewed • Not reviewed- Deleted: Recommendation from the previous CPG that has been removed because it was deemed out of scope for the updated CPG
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The VA/DOD guideline cited 2 RCTs, 2 systematic reviews, and 1 clinical study. In the VA/DOD guideline, several of the same RCTs were cited as the APA and ICSI guidelines. One of the systematic review (Vergouwen et al., 2003) examined antidepressant medication adherence, and found that collaborative care approaches consistently enhanced adherence during both the acute and continuation phase of treatments, as well as led to improved clinical benefits.
Estimates of benefit and consistency across studies	The benefit of adherence to antidepressant medications is a reduction in the recurrence rate of new episodes of depression. The guidelines and evidence note that pharmacotherapy is most effective when the physician identifies the most effective medication for each patient. This performance measure focuses on continuation of medication during the acute and continuation phases of treatment. Evidence suggests that physicians can help maximize the efficacy of medication treatment by monitoring the effects of the medication and the dosage. Across included studies, guidelines agree that antidepressants are an effective way to treat people with major depression, if steps are taken to help patients adhere to their medications.

What harms were identified?	The guidelines and evidence note that pharmacotherapy is most effective when the physician identifies the most effective medication for each patient. The harms stem from a lack of adherence to medications.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 3: Clinical Practice Guideline 3

Source of Clinical Practice Guideline: o **Title** o **Author** o **Date** o **Citation, including page number** o **URL**	o Institute for Clinical Systems Improvement. Major Depression in Adults in Primary Care o Trangle, M., et al. o May 2012 o Trangle M, Dieperink B, Gabert T, Haight B, Lindvall B, Mitchell J, Novak H, Rich D, Rossmiller D, Setter¬lund L, Somers K. Institute for Clinical Systems Improvement. Major Depression in Adults in Primary Care.http://bit.ly/Depr0512. Updated May 2012. o Institute for Clinical Systems Improvement: Recommendations for the Diagnosis and Treatment of Major Depression in Adults in Primary Care o Trangle, M., et al. o March 2016 o Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, Mack N, Myszkowski M. Adult depression in primary care. Bloomington (MN): Institute for Clinical Systems Improvement (ICSI); 2016 Mar. 131 p. [394 references] o https://guideline.gov/summaries/summary/50406/adult-depression-in-primary-care?q=Depression+Adult+in+Primary+Care
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.	1. "For patients with chronic major depression, start with combined antidepressant medication and psychotherapy." (Quality of Evidence: High; Strength of Recommendation: Strong) •"Antidepressant medications and/or referral for psychotherapy are recommended as treatment for major depression. Factors to consider in making treatment recommendations are symptom severity, presence of psychosocial stressors, presence of comorbid conditions, and patient preferences. Physical activity and active patient engagement are also useful in easing symptoms of major depression. 2. "Before initiating treatment, it is important to establish a therapeutic alliance with the patient regarding diagnosis and treatment options

	(in which there is overlap in the patient's and clinician's definition of the problem and agreement on which steps are to be taken by each)." (Quality of Evidence: Low; Strength of Recommendation: Strong) 3. "Clinicians should establish and maintain follow-up with patients." (Quality of Evidence: High; Strength of Recommendation: Strong) •If the primary care provider is seeing incremental improvement, continue working with that patient to increase medication dosage or augment with psychotherapy or medication to reach remission. This can take up to three months. Don't give up on the patient whether treating in primary care or referring. Studies have shown that primary care can be just as successful as specialty care. - For medication treatment, patients may show improvement at two weeks but need a longer length of time to really see response and remission. Most people treated for initial depression need to be on medication at least 6-12 months after adequate response to symptoms. Patients with recurrent depression need to be treated for three years or more."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	~~"Guideline" grade~~ Evidence is reviewed using Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology. The work group then reaches consensus and categorizes evidence into the following categories for use in the guideline: High: Further research is very unlikely to change confidence in the estimate of effect. Low: Further research is very likely to have an important impact on confidence in the estimate of effect and is likely to change the estimate or any estimate of effect is very uncertain.
Provide all other grades and definitions from the evidence grading system	Moderate Quality Evidence: Further research is likely to have an important impact on confidence in the estimate of effect and may change the estimate.
Grade assigned to the **recommendation** with definition of the grade	1 and 3: High Quality Evidence with Strong Recommendation: The work group is confident that the desirable effects of adhering to this recommendation outweigh the undesirable effects. This is a strong recommendation for or against. This applies to most patients. 2: Low Quality Evidence with Strong Recommendation: The work group feels that the evidence consistently indicates the benefit of this action outweighs the harms. This recommendation might change when higher quality evidence becomes available.
Provide all other grades and definitions from the recommendation grading system	~~ICSI RATING SCHEME FOR THE STRENGTH OF THE RECOMMENDATION~~ ~~GRADE Methodology~~ High Quality Evidence with Weak Recommendation: The work group recognizes that the evidence, though of high quality, shows a balance between estimates of harms and benefits. The best action will depend on local circumstances, patient values or preferences.

	Moderate Quality Evidence with Strong Recommendation: The work group is confident that the benefits outweigh the risks, but recognizes that the evidence has limitations. Further evidence may impact this recommendation. This is a recommendation that likely applies to most patients. Moderate Quality Evidence with Weak Recommendation: The work group recognizes that there is a balance between harms and benefit, based on moderate quality evidence, or that there is uncertainty about the estimates of the harms and benefits of the proposed intervention that may be affected by new evidence. Alternative approaches will likely be better for some patients under some circumstances. Low Quality Evidence with Weak Recommendation: The work group recognizes that there is significant uncertainty about the best estimates of benefits and harms.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Frequently refers to APA guideline. The ICSI guideline includes 3 studies showing high level evidence (GRADE rating), 1 systematic review, and 5 studies showing low level evidence (GRADE rating), The three studies demonstrating high level evidence were RCTs that looked at the impact of adherence on relapse with various numbers of participants (386, 153, and 386 respectively). One of the five studies showing low level of evidence was an observational study that looked at a total of 4,052 patients with major depression and the effect of antidepressant maintenance on relapse rates.
Estimates of benefit and consistency across studies	With regards to initiating treatment, the cited evidence found consistency across studies that "antidepressant treatment with psychotherapy outperforms either treatment as monotherapy and more rapidly begins the process of reversing symptoms, suffering and functional impairment in a condition that can go on for decades untreated. Psychotherapy can produce quality-of-life improvements and lower health and human services costs." With regards to follow-up with patients in treatment, the cited evidence found consistency across studies that "appropriate, reliable follow-up is highly correlated with improved response and remission scores. It is also correlated with the improved safety and efficacy of medications and helps prevent relapse."
What harms were identified?	With regards to initiating treatment, "Combined medication and psychotherapy increase short-term costs. Access to high-quality psychotherapy is not available in many primary care settings. In a 2000 study of chronic major depression, which excluded pure dysthymic disorder, the overall drop-out rate was the same for the three treatment groups, but reasons for dropping out varied. More patients dropped out of the medication-alone arm because of adverse events, and more psychotherapy patients withdrew consent because therapy was too time consuming, they did not want psychotherapy, or they wanted medication. This highlights the need to consider patient preferences. The benefits of psychotherapy are delayed and may cause some patients to give up on it prematurely."

	Potential harms associated with proper follow-up care with patients in treatment may include added expense and unnecessary visits. However, "Benefits appear to outweigh potential harms by a wide margin."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

Table 4: Meta-Analysis 1

Source of Meta-Analysis: o **Title** o **Author** o **Date** o **Citation, including page number** o **URL**	• Antidepressant Drug effects and Depression Severity: A Patient Level Meta-Analysis • Fournier, J., et al. • January, 2010 • Fournier, J.C., et al. 2010. Antidepressant drug effects and depression severity: A patient-level meta-analysis. JAMA 303(1): 47-53. • https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3712503/pdf/nihms483345.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.	This meta-analysis concluded that, among studies comparing the relative benefit of antidepressant medication vs placebo in the treatment of major or minor depressive disorder, the magnitude of superiority increased as baseline severity of depression increased, as measured by the Hamilton Rating Scale for Depression.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The evidence was not graded
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	The evidence was not graded
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The sample consisted of participants from five randomized placebo-controlled trials of an FDA approved antidepressant in the treatment of Major or Minor Depressive Disorder (five major depressive disorder, one minor depression). The pooled sample for the analysis included 434 patients in the antidepressant medication (ADM) group and 284 patients in the placebo group.
Estimates of benefit and consistency across studies	Across the data from the included studies, this meta-analysis found "the efficacy of ADM treatment for depression varies considerably as a function of symptom severity." The results suggest that for mild and moderate depression baseline symptoms, ADM treatment may

	not demonstrate significant results when compared to placebo. The study builds off earlier work by Zimmerman et al., that suggests Hamilton Depression Rating Scale scores of 18-20 are appropriate for ADM, and instead finds that baseline scores of 25 and over show a significant ADM drug-placebo difference.
What harms were identified?	No harms were identified. Patients who initiated treatment with ADM with baseline scores below 25 demonstrated nonexistent-to-negligible drug effects.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	None identified.

Table 5: Meta-Analysis 2

Source of Meta-Analysis: o **Title** o **Author** o **Date** o **Citation, including page number** o **URL**	• Antidepressants for treatment of depression in primary care: a systematic review and meta-analysis. • Arroll, B., et al. • December 2016. • Arroll, B., et al. 2016. Antidepressants for treatment of depression in primary care: a systematic review and meta-analysis. J Prim Health Care. 8(4): 325-334. • https://www.ncbi.nlm.nih.gov/pubmed/29530157
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.	"This study updates the Cochrane review by including newer antidepressant classes and calculating numbers needed to treat (NNTs) for individual drugs where data were available. There was evidence to support the effectiveness of tricyclic antidepressants (TCAs) and serotonin selective reuptake inhibitors (SSRIs) when compared to placebo, and evidence of efficacy for serotonin–norepinephrine reuptake inhibitor (SNRIs) and noradrenergic and specific serotonergic antidepressant (NaSSA)."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The evidence was not graded
Provide all other grades and definitions from the evidence grading system	
Grade assigned to the **recommendation** with definition of the grade	The evidence was not graded
Provide all other grades and definitions from the recommendation grading system	
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The final review included 17 randomized control trials. Selection criteria included antidepressant studies with a randomly assigned placebo group where half or more subjects were recruited from primary care.
Estimates of benefit and consistency across studies	The authors discuss consistency across medication-to-placebo studies that conclude antidepressants are effective for patients in primary care with depression.
What harms were identified?	No harms were identified.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

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.

~~Studies found outside guidelines: 2 RCTs; 1 systematic review; 1 meta-analysis; 2 fact sheets; 1 qualitative review; 2 prospective studies; 1 survey study; 1 case study. One of the RCT study (Rost, 2001) looked at 479 adult patients from 12 primary care practices to identify primary care practices that improved adherence to medication for new episodes of depression. The referenced meta-analysis (Fournier, 2010) identified randomized placebo-controlled trials that examined whether antidepressant medication represented effective treatment for people with major depression and found substantial evidence to support pharmacotherapy.~~

Studies found outside guidelines: 2 RCTs; 1 qualitative reviews; 2 prospective studies; 2 survey studies; 1 case study; and 1 fact sheet. One RCT study (Rost, 2001) looked at 479 adult patients from 12 primary care practices to identify primary care practices that improved adherence to medication for new episodes of depression.

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

The body of evidence found that the use and adherence of antidepressants were associated with better outcomes for people in terms of lower rates of relapse and lower rates of new episodes of major depression. The evidence and the focus of this measure, adherence to antidepressants for people with major depression, are directly related.

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

A targeted literature review was conducted to identify evidence.

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

Kessler RC, Chiu WT, Demler O, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62:617–627

Burcusa, S.L., W.G. Iacono. 2007. Risk for recurrence in depression. Clin Psychol Rev 27(8): 959-85.

Melartin, T.K., H.J. Rytsala, U.S. Leskela, P.S. Lestela-Mielonen, T.P. Sokero, E.T. Isometsa. 2005. Continuity is the main challenge in treating major depressive disorder in psychiatric care. J Clin Psychiatry 66(2):220-7.

Johnston, K., W. Westerfield, S. Momim, R. Phillipi. 2009. The direct and indirect costs of employee depression, anxiety, and emotional disorders—An employer case study. J of Occ and Envt Med 51(5): 564-77.

Katon W, Russo, J, Von Korff M, et al. Long-term effects of a collaborative care intervention in persistently depressed primary care patients. J Gen Intern Med. 2002;17:741-748.

Rost K, Nutting P, Smith J, et al. Improving depression outcomes in the community primary care practice: a randomized trial of the quest intervention. Quality Enhancement by Strategic Teaming. J Gen Intern Med. 2001;16:143-149.

Simon, G.E. 2002. Evidence review: efficacy and effectiveness of antidepressant treatment in primary care. Gen Hosp Psychiatry 24(4):213-24.

Stewart, W.F., J.A. Ricci, E. Chee, S.R. Hahn, D. Morganstein. 2003. Cost of lost productive work time among US workers with depression. JAMA 289(23):3135-44.

The National Alliance on Mental Illness. 2009. Major Depression Fact Sheet. http://www.nami.org/Template.cfm?Section=Depression&Template=/ContentManagement/ContentDisplay.cfm&ContentID=88956 (October 27, 2011)

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 0105
Corresponding Measures:
De.2. **Measure Title:** Antidepressant Medication Management (AMM)
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of members 18 years of age and older who were treated antidepressant medication, had a diagnosis of major depression, and who remained on an antidepressant medication treatment. Two rates are reported.

a) Effective Acute Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 84 days (12 weeks).
b) Effective Continuation Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 180 days (6 months).

a) Effective Acute Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 84 days (12 weeks).
b) Effective Continuation Phase Treatment. The percentage of patients who remained on an antidepressant medication for at least 180 days (6 months).
1b.1. **Developer Rationale:** Clinical guidelines for depression emphasize the importance of effective clinical management in increasing patients' medication compliance, monitoring treatment effectiveness, and identifying and managing side effects. If pharmacological treatment is initiated, appropriate dosing and continuation of therapy through the acute and continuation phases decrease recurrence of depression. Thus, evaluation of duration of pharmacological treatment serves as an important indicator in promoting patient compliance with the establishment and maintenance of an effective medication regimen.

S.4. **Numerator Statement:** Adults 18 years of age and older who were newly treated with antidepressant medication, had a diagnosis of major depression, and who remained on an antidepressant medication treatment.
S.6. **Denominator Statement:** Patients 18 years of age and older with a diagnosis of major depression and were newly treated with antidepressant medication.
S.8. **Denominator Exclusions:** Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients who did not have a diagnosis of major depression in an inpatient, outpatient, ED, telehealth, intensive outpatient or partial hospitalization setting during the 121-day period from 60 days prior to the IPSD, through the IPSD and the 60 days after the IPSD.

Exclude patients who filled a prescription for an antidepressant 105 days prior to the IPSD.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Feb 28, 2014

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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
0105._evidence_attachment_7.1_FINAL.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.
Yes

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 for depression emphasize the importance of effective clinical management in increasing patients' medication compliance, monitoring treatment effectiveness, and identifying and managing side effects. If pharmacological treatment is initiated, appropriate dosing and continuation of therapy through the acute and continuation phases decrease recurrence of depression. Thus, evaluation of duration of pharmacological treatment serves as an important indicator in promoting patient compliance with the establishment and maintenance of an effective medication regimen.

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 following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. It includes the number of health plans and the average eligible population across plans for which the measure applies. Performance data is summarized at the health plan level. Performance of health plans is represented by percentiles, inter-quartile range, mean, min, max and standard deviations. Data is stratified by year and product line (i.e. commercial, Medicare, Medicaid).

Commercial – Effective Acute Phase Treatment
Measurement Year: 2011; 2010; 2009
Number of Plans: 385; 398; 414
Mean Eligible Population: 864; 828; 819
Mean: 65.3; 64.6; 63
Standard Deviation: 6.23; 6.48; 6.57
Standard Error: 0.32; 0.32; 0.32
Minimum: 38.9; 35.5; 31.7
Maximum: 85.5; 90.4; 89
P10: 58; 57.4; 55.9
P25: 61.5; 60.7; 59.1
P50: 64.8; 64.8; 63

P75: 69.1; 68.1; 66.6
P90: 72.3; 72.2;70.8

Commercial – Effective Acute Phase Treatment
Measurement Year: 2017; 2016; 2015
Number of Plans: 403; 407; 398
Mean Eligible Population: 1,958; 1,927; 1,939
Mean: 67.5; 66.5; 66.1
Standard Deviation: 6.5; 6.9; 6.9
Standard Error: 0.32; 0.34; 0.35
Minimum: 39.1; 27.0; 30.7
Maximum: 84.4; 85.1; 82.7
P10: 58.6; 58.6; 58.1
P25: 64.0; 62.7; 62.0
P50: 67.5; 66.6; 65.7
P75: 71.8; 71.0; 71.0
P90: 75.7; 74.3; 75.2

Commercial – Effective Continuation Phase Treatment
Measurement Year: 2011; 2010; 2009
Number of Plans: 385; 398; 414
Mean Eligible Population: 864; 828; 819
Mean: 49.1; 48.2; 46.3
Standard Deviation: 6.57; 6.95; 7.12
Standard Error: 0.33; 0.35; 0.35
Minimum: 27;19.4; 15.8
Maximum: 76.6; 87.2; 77
P10: 41.9; 39.8; 38.2
P25: 44.8; 44.2; 42
P50: 48.9; 48.2; 45.7
P75: 53.3; 52.3; 50
P90: 56.9; 55.7; 54.5

Commercial – Effective Continuation Phase Treatment
Measurement Year: 2017; 2016; 2015
Number of Plans: 403; 407; 398
Mean Eligible Population: 1,958; 1,927; 1,939
Mean: 51.8; 50.7; 50.3
Standard Deviation: 6.8; 7.1; 7.4
Standard Error: 0.34; 0.35; 0.37
Minimum: 21.9; 18.9; 22.7
Maximum: 70.4; 75.9; 75.0
P10: 43.4; 42.6; 42.0
P25: 47.6; 46.7; 45.7
P50: 51.5; 50.5; 49.8
P75: 56.0; 55.3; 54.6
P90: 60.4; 58.8; 59.9

Medicaid – Effective Acute Phase Treatment
Measurement Year: 2011; 2010; 2009
Number of Plans: 97; 90; 76
Mean Eligible Population: 505; 493; 380
Mean: 51.1; 50.7; 49.7
Standard Deviation: 7.7; 8.16; 8.69
Standard Error: 0.78; 0.86; 1
Minimum: 37.5; 30; 30.2
Maximum: 81; 78.9; 84.7

P10: 43.4; 43; 40.9
P25: 47; 46.4; 45.2
P50: 49.4; 50.1; 48.1
P75: 52.7; 53.6; 53.2
P90: 61.6; 59.9; 58.4

Medicaid – Effective Acute Phase Treatment
Measurement Year: 2017; 2016; 2015
Number of Plans: 226; 216; 188
Mean Eligible Population: 2,301; 1,855; 1,377
Mean: 53.2; 54.5; 52.4
Standard Deviation: 8.9; 9.9; 9.6
Standard Error: 0.59; 0.67; 0.70
Minimum: 17.1; 23.6; 17.7
Maximum: 99.1; 94.8; 92.3
P10: 44.5; 44.0; 42.8
P25: 48.2; 48.4; 46.7
P50: 51.9; 53.5; 50.5
P75: 57.5; 60.0; 56.3
P90: 64.2; 67.6; 62.7

Medicaid – Effective Continuation Phase Treatment
Measurement Year: 2011; 2010; 2009
Number of Plans: 97;90; 76
Mean Eligible Population: 505; 493; 380
Mean: 34.4; 34.4; 33
Standard Deviation: 7.91; 9.11; 9.86
Standard Error: 0.8; 0.96; 1.13
Minimum: 20.4; 17.6; 12.5
Maximum: 67.1; 74.6; 80.5
P10: 26.7; 25.7; 24.8
P25: 30; 29.2; 27.8
P50: 32.4; 32.7; 31
P75: 37.3; 37.5; 35.4
P90: 42.9; 44.2; 43.3

Medicaid – Effective Continuation Phase Treatment
Measurement Year: 2017; 2016; 2015
Number of Plans: 226; 218; 188
Mean Eligible Population: 2,301; 1,855; 1,377
Mean: 38.0; 39.5; 37.1
Standard Deviation: 9.4; 10.6; 10.6
Standard Error: 0.63; 0.72; 0.77
Minimum: 8.6; 11.5; 8.8
Maximum: 82.3; 84.2; 88.8
P10: 29.1; 28.1; 27.4
P25: 32.6; 32.8; 30.9
P50: 36.3; 38.1; 34.0
P75: 41.6; 43.5; 40.8
P90: 50.4; 54.3; 49.8

Medicare – Effective Acute Phase Treatment
Measurement Year: 2011; 2010; 2009
Number of Plans: 335; 278; 241
Mean Eligible Population: 220; 195; 186
Mean: 67.6; 65.6; 63.6
Standard Deviation: 10.4; 10.5; 10.8

Standard Error: 0.57; 0.63; 0.7
Minimum: 33.3; 25.9; 25.5
Maximum: 94.7; 92.8; 93.5
P10: 52.4; 53.5; 50.8
P25: 62.2; 59.3; 57.5
P50: 68.4; 65.6; 63.8
P75: 74.3; 72.4; 70.1
P90: 79.6; 77.6; 76.3

Medicare – Effective Acute Phase Treatment
Measurement Year: 2017; 2016; 2015
Number of Plans: 401; 384; 388
Mean Eligible Population: 1,010; 943; 807
Mean: 70.2; 70.1; 69.4
Standard Deviation: 8.8; 9.8; 8.8
Standard Error: 0.44; 0.50; 0.45
Minimum: 38.7; 13.9; 38.6
Maximum: 99.2; 100.0; 90.9
P10: 59.4; 58.1; 57.8
P25: 64.8; 64.6; 64.0
P50: 70.7; 70.3; 70.0
P75: 76.0; 76.1; 75.4
P90: 80.3; 82.5; 79.2

Medicare – Effective Continuation Phase Treatment
Measurement Year: 2011; 2010; 2009
Number of Plans: 335; 278; 241
Mean Eligible Population: 220; 195; 186
Mean: 54.8; 52.8; 50.6
Standard Deviation: 11.3; 11.4; 11.7
Standard Error: 0.62; 0.68; 0.75
Minimum: 20.2; 14.1; 16.9
Maximum: 89.4; 84.5; 87.1
P10: 39.1; 38.1; 36.9
P25: 48.5; 46.4; 43.5
P50: 55.8; 53; 50.9
P75: 62.4; 60.7; 57.1
P90: 68.2; 66.1; 65.8

Medicare – Effective Continuation Phase Treatment
Measurement Year: 2017; 2016; 2015
Number of Plans: 401; 384; 388
Mean Eligible Population: 1,010; 943; 807
Mean: 55.5; 56.2; 55.7
Standard Deviation: 10.3; 11.3; 10.3
Standard Error: 0.52; 0.58; 0.52
Minimum: 8.6; 11.5; 8.8
Maximum: 93.0; 96.7; 86.1
P10: 42.1; 42.9; 42.6
P25: 48.9; 49.5; 49.2
P50: 55.6; 56.2; 55.9
P75: 61.2; 61.9; 62.4
P90: 67.5; 70.3; 68.5

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.

The data in 1b.2 are HEDIS health plan performance rates.

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.*

• The percentage of adults with a major depressive episode in 2008, who received treatment for it, was significantly lower for blacks than for whites (58.9 vs. 71.1 percent) and for Hispanics than non-Hispanic whites (51.8 vs. 73.3 percent). (AHRQ, 2009)
• A study examining antidepressant treatment patterns found that, compared to younger adults, older adults tended to be more likely to discontinue antidepressant treatment (Sanglier et al., 2011).
• A study that examined the treatment disparities for respondents with major depressive disorders showed that blacks and Hispanics were less likely to use antidepressants than whites. Of the respondents who were screened, only 34% reported antidepressant use in the previous 12-month period; however, blacks (17.5%) and Hispanics (21.8%) reported statistically significant lower overall use of antidepressants in analysis compared with whites (37.6%) (Fleming et al., 2003).
• Compared to whites, blacks and Hispanics in primary care were less likely to be prescribed antidepressants for their depression. Whites also received more antidepressant prescriptions after a visit to psychiatrists when compared to blacks (Lagomasino et al., 2011).

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**

Agency for Healthcare Research and Quality. Mental Health Research Findings. Program Brief. AHRQ Publication No. 09-P011, September 2009. Rockville, MD. http://www.ahrq.gov/research/mentalhth.htm

Sanglier T, Saragoussi D, Milea D, Auray JP, Valuck RJ, Tournier M., Comparing antidepressant treatment patterns in older and younger adults: a claims database analysis. J Am Geriatr Soc. 2011 Jul;59(7):1197-205. doi: 10.1111/j.1532-5415.2011.03457.x. Epub 2011 Jun 30. http://www.ncbi.nlm.nih.gov/pubmed/21718261

Fleming M, Barner JC, Brown CM, Smith T. Treatment disparities for major depressive disorder: Implications for pharmacists.J Am Pharm Assoc. 2003. 2011 Sep-Oct;51(5):605-12.

Lagomasino IT, Stockdale SE, Miranda J. Racial-ethnic composition of provider practices and disparities in treatment of depression and anxiety, 2003-2007. Psychiatr Serv. 2011 Sep;62(9):1019-25

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):*
Behavioral Health, Behavioral Health : Depression

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

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

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 Applicable

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:** 0105_AMM_Value_Sets_updated_4.11.18.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.
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.
No important changes have been made to the measure since the last update.

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).
Adults 18 years of age and older who were newly treated with antidepressant medication, had a diagnosis of major depression, and who remained on an antidepressant medication treatment.

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).
a) Effective Acute Phase Treatment: At least 84 days (12 weeks) of treatment with antidepressant medication (Table AMM-C) during the 114-day period following the Index Prescription Start Date (IPSD) (115 total days). This allows gaps in medication treatment up to a total of 31 days during the 115-day period. Gaps can include either washout period gaps to change medication or treatment gaps to refill the same medication.

b) Effective Continuation Phase Treatment: At least 180 days (6 months) of continuous treatment with antidepressant medication (Table AMM-C) during the 231-day period following the IPSD (232 total days). This allows gaps in medication treatment up to a total of 52 days during the 232-day period. Gaps can include either washout period gaps to change medication or treatment gaps to refill the same medication.

TABLE AMM-C: ANTIDEPRESSANT MEDICATIONS
Miscellaneous antidepressants: Bupropion, Vilazodone, Vortioxetine

Monoamine oxidase inhibitors: Isocarboxazid, Phenelzine, Selegiline, Tranylcypromine

Phenylpiperazine antidepressants: Nefazodone, Trazodone

Psychotherapeutic combinations: Amitriptyline-chlordiazepoxide, Amitriptyline-perphenazine, Fluoxetine-olanzapine

SNRI antidepressants : Desvenlafaxine, Duloxetine, Levomilnacipran, Venlafaxine

SSRI antidepressants: Citalopram, Escitalopram, Fluoxetine, Fluvoxamine, Paroxetine, Sertraline

Tetracyclic antidepressants: Maprotiline, Mirtazapine

Tricyclic antidepressants: Amitriptyline, Amoxapine, Clomipramine, Desipramine, Doxepin (>6mg), Imipramine, Nortriptyline, Protriptyline, Trimipramine

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18 years of age and older with a diagnosis of major depression and were newly treated with antidepressant medication.

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).
Step 1: Determine the Index Prescription Start Date (IPSD). Identify the date of the earliest dispensing event for an antidepressant medication (Table AMM-C) during the Intake Period (The 12-month window starting on May 1 of the year prior to the measurement year and ending on April 30 of the measurement year).

Step 2: Required exclusion: Exclude patients who did not have a diagnosis of major depression in an inpatient, outpatient, ED, telehealth, intensive outpatient or partial hospitalization setting during the 121-day period from 60 days prior to the IPSD, through the IPSD and the 60 days after the IPSD. Patients who meet any of the following criteria remain in the eligible population:

• An outpatient visit, ED visit, telehealth, intensive outpatient encounter or partial hospitalization with any diagnosis of major depression. Either of the following code combinations meets criteria:
– AMM Stand Alone Visits Value Set with Major Depression Value Set. with or without a telehealth modifier (Telehealth Modifier Value Set).
– AMM Visits Value Set with AMM POS Value Set and Major Depression Value Set, with or without a telehealth modifier (Telehealth Modifier Value Set). .
• Telephone Visits Value Set with Major Depression Value Set.
• An ED visit (ED Value Set) with any diagnosis of major depression (Major Depression Value Set).
• An acute or nonacute inpatient stay discharge with any diagnosis of major depression (Major Depression Value Set). To identify acute and nonacute inpatient discharges:
First, identify all acute and nonacute inpatient stays (Inpatient Stay Value Set). Second, identify the admission and discharge dates for the stay. Either an admission or discharge during the required time frame meets criteria.

Step 3: Test for Negative Medication History. Exclude patients who filled a prescription for an antidepressant medication 105 days prior to the IPSD.

Step 4: Calculate continuous enrollment. Patients must be continuously enrolled for 105 days prior to the IPSD to 231 days after the IPSD.

TABLE AMM-C: ANTIDEPRESSANT MEDICATIONS
Miscellaneous antidepressants: Bupropion, Vilazodone, Vortioxetine

Monoamine oxidase inhibitors: Isocarboxazid, Phenelzine, Selegiline, Tranylcypromine

Phenylpiperazine antidepressants: Nefazodone, Trazodone

Psychotherapeutic combinations: Amitriptyline-chlordiazepoxide, Amitriptyline-perphenazine, Fluoxetine-olanzapine

SNRI antidepressants : Desvenlafaxine, Duloxetine, Levomilnacipran, Venlafaxine

SSRI antidepressants: Citalopram, Escitalopram, Fluoxetine, Fluvoxamine, Paroxetine, Sertraline

Tetracyclic antidepressants: Maprotiline, Mirtazapine

Tricyclic antidepressants: Amitriptyline, Amoxapine, Clomipramine, Desipramine, Doxepin (>6mg), Imipramine, Nortriptyline, Protriptyline, Trimipramine

*See corresponding Excel file for value sets referenced above.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients who did not have a diagnosis of major depression in an inpatient, outpatient, ED, telehealth, intensive outpatient or partial hospitalization setting during the 121-day period from 60 days prior to the IPSD, through the IPSD and the 60 days after the IPSD.

Exclude patients who filled a prescription for an antidepressant 105 days prior to the IPSD.

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.)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These patients may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

Exclude patients who did not have a diagnosis of major depression in an inpatient, outpatient, ED, telehealth, intensive outpatient or partial hospitalization setting during the 121-day period from 60 days prior to the IPSD, through the IPSD and the 60 days after the IPSD. Patients who meet any of the following criteria remain in the eligible population:

• An outpatient visit, ED visit, telehealth, intensive outpatient encounter or partial hospitalization with any diagnosis of major depression. Either of the following code combinations meets criteria:
– AMM Stand Alone Visits Value Set with Major Depression Value Set, with or without a telehealth modifier (Telehealth Modifier Value Set).

– AMM Visits Value Set with AMM POS Value Set and Major Depression Value Set, with or without a telehealth modifier (Telehealth Modifier Value Set).
• Telephone Visits Value Set with Major Depression Value Set.
• An ED visit (ED Value Set) with any diagnosis of major depression (Major Depression Value Set).

• An acute or nonacute inpatient stay with any diagnosis of major depression (Major Depression Value Set). To identify acute and nonacute inpatient discharges:

First, identify all acute and nonacute inpatient stays (Inpatient Stay Value Set). Second, identify the admission and discharge dates for the stay. Either an admission or discharge during the required time frame meets criteria.

Exclude patients who filled a prescription for an antidepressant medication 105 days prior to the IPSD.

*See corresponding Excel file for value sets referenced above.

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.)*
NCQA asks that health plans collect the measure data for each of the three product lines each year (i.e. commercial, Medicare, Medicaid) if applicable.

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: Determine the eligible population, or denominator.
Step 1a: Determine the Index Prescription Start Date (IPSD). Identify the date of the earliest dispensing event for an antidepressant medication (Table AMM-C) during the Intake Period (the 12-month window starting on May 1 of the year prior to the measurement year and ending on April 30 of the measurement year).
Step 1b: Exclude patients who did not have a diagnosis of major depression in an inpatient, outpatient, ED, telehealth, intensive outpatient or partial hospitalization setting during the 121-day period from 60 days prior to the IPSD, through the IPSD and the 60 days after the IPSD.

Step 1c: Test for Negative Medication History. Exclude patients who filled a prescription for an antidepressant medication 105 days prior to the IPSD.
Step 1d: Calculate continuous enrollment. Exclude patients who are not continuously enrolled for 105 days prior to the IPSD to 231 days after the IPSD.

Step 2: Determine the numerators for the two reported rates.
Step 2a (Effective Acute Phase Treatment): Identify at least 84 days (12 weeks) of continuous treatment with antidepressant medication (Table AMM-C) during the 114-day period following the Index Prescription Start Date (IPSD) (115 total days). This allows gaps in medication treatment up to a total of 31 days during the 115-day period. Gaps can include either washout period gaps to change medication or treatment gaps to refill the same medication.
Step 2b (Effective Continuation Phase Treatment): Identify at least 180 days (6 months) of continuous treatment with antidepressant medication (Table AMM-C) during the 232-day period following the IPSD. Continuous treatment allows gaps in medication treatment up to a total of 52 days during the 232-day period. Gaps can include either washout period gaps to change medication or treatment gaps to refill the same medication.
Step 3: Calculate the two reported rates by dividing both the numerators from steps 2a and 2b by the denominator in step 1d.

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.
N/A

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

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

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)*
Health Plan

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.)*

2. Validity – See attached Measure Testing Submission Form
0105_-_Antidepressant_Medication_Management_-_Testing_Form_v7.1_FINAL.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 0105
Measure Title: Antidepressant Medication Management
Date of Submission: 4/2/2018
Type of Measure:

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

Instructions

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

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

2018 submission

2012 Submission

This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via the Interactive Data Submission System (IDSS) portal.
The URL is: http://www.ncqa.org/tabid/370/default.aspx

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

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

Measure Specified to Measure Performance of:	**Measure Tested at Level of:**

(*must be consistent with levels entered in item S.20*)	
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other:	☐ other: Click here to describe

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

2018 Submission

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

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

2012 Submission

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

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

2018 Submission

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

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

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

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

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

2018 Submission

N/A

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

2018 Submission

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

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

2018 Submission

Reliability Testing of Performance Measure Score: same as below

2012 submission

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

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies

that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

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

2018 Submission

Beta-binomial statistic for each measure rate:

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

2012 submission

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

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

2018 Submission

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

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

2018 submission:

We assessed both construct and face validity for this measure.

Method of testing construct validity:

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

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

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

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

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

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

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

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

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

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

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

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

2012 submission:

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

NCQA uses a standardized process called the HEDIS measure life cycle to ensure the validity of measures.

*Step 1: Topic selection is the process of identifying measures that meet criteria consistent with the overall model for performance measurement. There is a huge universe of potential performance measures for future versions of HEDIS. The first step is identifying measures that meet formal criteria for further development.

NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted.

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

The Committee on Performance Measurement (CPM) uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

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

*Step 4: First-year data collection requires that organizations collect and report first-year measures and that those measures be available for audit. First-year measure results are not publicly reported and are not included in NCQA's Quality Compass or in accreditation scoring.

After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

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

Step 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reevaluated at least every three years.

AMM MEASURE DEVELOPMENT AND TESTING:

Step 1: NCQA developed the Antidepressant Medication Management measure to address the gap in care surrounding adherence to antidepressants for people diagnosed with major depression. NCQA's Performance Measurement Department and the Behavioral Health MAP worked together to assess the most appropriate elements of this measure.

Step 2: The measure was written, field-tested, and presented to the CPM and incorporated into HEDIS in 1998 for HEDIS 1999. After reviewing field test results, The CPM's recommendation was to send the measure to public comment with a majority vote.

Step 3: NCQA released the measure for Public Comment prior to publication in HEDIS. We received and responded to comments on this measure. Based on positive feedback, the CPM recommended moving this measure to first year data collection by a majority vote.

Step 4: The Antidepressant Medication Management measure was introduced in HEDIS 1999. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure public reporting with a majority vote.

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

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

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

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

2018 Submission

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

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

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

Note: $p<0.0001$

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

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

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

Note: $p<0.0001$

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

Pearson Correlation Coefficients	Statin Therapy for Patients With Diabetes (Statin Adherence Indicator: Members who remained on a statin medication of any intensity for at least

	80% of the treatment period)
Antidepressant Medication Management – Acute Phase	0.56
Antidepressant Medication Management – Continuation Phase	0.60

Note: $p<0.0001$

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

2012 submission:
NCQA field tested the measure in 1998 and again in 2007.

NCQA developed the measure through the Robert Wood Johnson Chronic Disease Grant. The field testing in 1996 included two health plans. The field test design had two goals:
1. Find out if the measure should focus on appropriate dosing of antidepressants or adherence.
2. See a relationship between adherence and depression relapse.

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

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

Question 1.
- For the three plans, the average percent of the eligible population with a major depression diagnosis was 9.5 percent.

Question 2:
- Through both administrative claims and medical record data, plans can find the length of time prior to the index prescription date that a member was prescribed an antidepressant.

Question 3:
- Health plans were concerned that 90 days is not sufficient to identify people currently on an antidepressant. Those concerns contend that extending the period would more accurately exclude people being treated with antidepressants prior to the index prescription date. The current negative medication history is 90 days, which aligns with the continuous enrollment period of 90 days. If the negative medication history was increased, to address this concern, an additional 4 percent of the eligible population would be excluded.

NCQA's experts felt that it was unnecessary to exclude more patients, because most prescriptions for antidepressants are for 90 days or less. Therefore, the measure accurately excludes people that are not newly treated with antidepressants.

Question 4:

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

Question 5:

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

Question 6:

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

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

2018 Submission

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

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

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

2b2. EXCLUSIONS ANALYSIS

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

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

Testing was not performed for exclusions.

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

measure scores)
Testing was not performed for exclusions.

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

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

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

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)*

2018 Submission

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

2012 submission

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

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

2018 Submission

HEDIS 2017 Variation in Performance across Health Plans for Acute Phase

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Commercial	1,958	67.5	6.5	58.6	64.0	67.5	71.8	75.7	7.8	<0.001

Medicare	1,010	70.2	8.8	59.4	64.8	70.7	75.9	80.3	11.1	<0.001
Medicaid	2,301	53.2	8.8	44.5	48.2	51.9	57.5	64.2	9.3	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile. P-values are less than 0.05.

HEDIS 2017 Variation in Performance across Health Plans for Continuation Phase

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

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile. P-values are less than 0.05.

2012 submission
There has been slow and steady improvement in performance in commercial, Medicare and Medicaid product lines over the last six years. Rates have gradually increased across means and percentiles at about the same rate. In general, rates are higher for the acute phase than the continuation phase, and higher in Medicare. Over the last three years, the number of plans reporting in the Medicare and Medicaid product lines has increased (close to 100 plans for Medicare), and dropped slightly in commercial. The data illustrates continued gaps in performance.

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
2018 Submission
The difference between the 25th and 75th percentile is statistically significant for both rates in all product lines.

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

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

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

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2018 Submission

This measure is collected with a complete sample.

2012 submission

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

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

handling missing data that were considered and pros and cons of each)

2018 Submission

This measure is collected with a complete sample.

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

2018 Submission

This measure is collected with a complete sample.

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)
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).
N/A

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

NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO's ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:

1) Information practices and control procedures

2) Sampling methods and procedures

3) Data integrity

4) Compliance with HEDIS specifications

5) Analytic file production

6) Reporting and documentation

In addition to the HEDIS audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.
Broad public use and dissemination of these measures are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Specific Plan for Use	Current Use (for current use provide URL)
Payment Program	Public Reporting Health Plan Rating http://www.ncqa.org/report-cards/health-plans/health-insurance-plan-ratings/ncqa-health-insurance-plan-ratings-2017 Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality

	Regulatory and Accreditation Programs NCQA Accreditation http://www.ncqa.org/tabid/123/Default.aspx Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality

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

MEDICAID ADULT CORE SET: There are a core set of health quality measures for Medicaid-enrolled adults. The Medicaid Adult Core Set was identified by the Centers of Medicare & Medicaid (CMS). The data collected from these measures helps CMS to better understand the quality of health care that adults enrolled in Medicaid receive nationally. Beginning in January 2014 and annually thereafter, the Secretary is required to publicly report the information that states voluntarily report to CMS on the quality of health care received by adults enrolled in Medicaid.

MERIT BASED INCENTIVE PAYMENT SYSTEM (MIPS) QUALITY PAYMENT PROGRAM (QPP): Eligible clinicians who elect to participate in MIPs earn a performance-based payment adjustment to Medicaid payments upon submission of evidence which attests that they provided high quality, efficient care supported by technology. Eligible clinicians can select up to six quality measures to report to CMS, including one outcome measure, that best fit their needs or specialty. The data collected from this program will help CMS to better understand the quality of health care that Medicare enrollees receive nationally.

HEALTH INSURANCE EXCHANGE QUALITY RATING SYSTEM (QRS): Qualified Health Plan (QHP) issuers and Multi-State Plan (MSP) issuers that offered coverage through a Health Insurance Marketplace (Marketplace) in the year prior to the current year are required to collect and submit QRS measure data to CMS. CMS produces quality ratings on a 5-star scale for each issuer in each State. Health plan level clinical quality measures and survey measures based on questions from the Qualified Health Plan Enrollee Experience Survey (QHP Enrollee Survey) are included in the QRS measure set. CMS collects data and calculates quality ratings for each QHP issuer's product type within each state and applies these ratings to each product type's QHPs in that State.

STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering a record 136 million people, or 43 percent of the U.S. population

HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported on the NCQA website. These ratings are based on a plan's performance on their HEDIS, CAHPS and accreditation standards scores. In 2017, a total of 521 Medicare Advantage health plans, 614 commercial health plans and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.

HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. As of Fall 2017, a total of 184 Medicare Advantage health plans were accredited using this measure among others covering 9.2 million Medicare beneficiaries; 451 commercial health plans covering 113 million lives; and 125 Medicaid health plans covering 35 million lives. Health plans are scored based on performance compared to benchmarks.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System.

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.
NCQA measures are evaluated regularly. During this "reevaluation" process, we seek broad input on the measure, including input on performance and implementation experience. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have not reported significant barriers to implementing this measure, as it uses the administrative data collection method. Questions have generally centered around minor clarification of the specifications, such as defining gaps in calculating days of medication treatment and questions about the supporting guidelines for the measure. NCQA responded to all questions to ensure consistent implementation of the specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the CMS Quality Rating System (QRS), CMS Merit-Based Incentive Payment System (MIPS) Program, and the Medicaid Adult Core Set.

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.**
Feedback has not required modification to this measure.

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.
Over the past three years, this measure has shown slight improvement (approximately an increase in performance by 1 percentage point across all product lines) across health plans (see section 1b.2 for summary of data from health plans). Of note, the highest performance continues to be seen in the Medicare population, for both the acute and continuation indicators. The Medicaid product continues to show the largest gap in performance, with performance consistently averaging about 17 percentage points lower than Medicare for both indicators.

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 were no identified unexpected findings during implementation of this measure.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits during implementation 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)**
#1880 – Adherence to Mood Stabilizers for People with Bipolar I Disorder.

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. Conceptually, these measures are similar, as the intent of both is to assess medication adherence for a specific population. #1880 is different from #0105 in two major ways: 1) it focuses on a population with bipolar disorder, rather than major depressive disorder, and 2) it tracks medication adherence using a “proportion of days covered” method, rather than a calculation of number of days of a dispensed prescription.

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

N/A

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.)
N/A

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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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 NCQA Behavioral Health Measurement Advisory Panel (BHMAP) is a balanced group of experts who have collaboratively advised NCQA throughout the development and maintenance of this measure. The BHMAP evaluated the measure specification at different stages of development and during reevaluations, reviewed field test results, and assessed NCQA's overall desirable attributes of relevance, scientific soundness, and feasibility. In addition to this advisory panel, this measure has been vetted with a host of other stakeholders, including our Committee on Performance Measurement (CPM), who voted on the measure for use in NCQA and related programs. All CPM recommendations are also reviewed and approved by NCQAs Board of Directors. Our measures are the result of consensus from a broad and diverse group of stakeholders.

Committee on Performance Measurement (CPM)

- Bruce Bagley, MD, American Academy of Family Physicians
- Andrew Baskin, MD, Aetna
- Jonathan Darer, MD, MPH, Medicalis
- Helen Darling, MA, City of Washington, DC
- Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
- Kate Goodrich, MD, MHS, Centers for Medicare & Medicaid Services
- David Grossman, MD, MPH, Kaiser Permanente Washington
- Christine S. Hunter, MD, US Office of Personnel Management
- Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
- Nancy Lane, PhD, Newton, MA
- Bernadette Loftus, MD, The Permanente Medical Group
- Adrienne Mims, MD, MPH, Alliant Quality
- Amanda Parsons, MD, MBA, Montefiore Health System
- Wayne Rawlins, MD, MBA, ConnectiCare

- Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
- Eric Schneider, MD, MSc, FACP, The Commonwealth Fund
- Marcus Thygeson, MD, MPH, San Rafael, CA
- JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms
- Lina Walker, PhD, AARP

Behavioral Health Measurement Advisory Panel:

- Katharine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute
- Christopher Dennis, MD, MBA, FAPA, Landmark Health, LLC
- Ben Druss, MD, MPH, Emory University
- Frank Ghinassi, PhD, ABPP, Rutgers University Behavioral Health Care
- Connie Horgan, ScD, Brandeis University
- Laura Jacobus-Kantor, PhD, SAMHSA
- Jeffrey Meyerhoff, MD, Optum
- Harold Pincus, MD, College of Physicians and Surgeons, Columbia University, New York Presbyterian Hospital, RAND
- Michael Schoenbaum, PhD, National Institute of Mental Health
- John Straus, MD, Massachusetts Behavioral Health Partnership-A Beacon Health Options Company

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 1998

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3 years, sooner if the clinical guidelines have changed significantly.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

Ad.6 **Copyright statement:** The performance measures and specifications were developed by and are owned by the National Committee for Quality Assurance ("NCQA"). The performance measures and specifications are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures and NCQA has no liability to anyone who relies on such measures or specifications. NCQA holds a copyright in these materials and can rescind or alter these materials at any time. These materials may not be modified by anyone other than NCQA. Anyone desiring to use or reproduce the materials without modification for an internal, quality improvement non-commercial purpose may do so without obtaining any approval from NCQA. All other uses, including a commercial use and/or external reproduction, distribution and publication must be approved by NCQA and are subject to a license at the discretion of NCQA.

Limited proprietary coding is contained in the measure specifications for convenience. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets. NCQA disclaims all liability for use or accuracy of any coding contained in the specifications.

Content reproduced with permission from HEDIS, Volume 2: Technical Specifications for Health Plans. To purchase copies of this publication, including the full measures and specifications, contact NCQA Customer Support at 888-275-7585 or visit www.ncqa.org/publications.

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 "AS IS" WITHOUT WARRANTY OF ANY KIND.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 1879
Measure Title: Adherence to Antipsychotic Medications for Individuals with Schizophrenia
Measure Steward: Centers for Medicare and Medicaid Services
Brief Description of Measure: Percentage of individuals at least 18 years of age as of the beginning of the measurement period with schizophrenia or schizoaffective disorder who had at least two prescription drug claims for antipsychotic medications and had a Proportion of Days Covered (PDC) of at least 0.8 for antipsychotic medications during the measurement period (12 consecutive months).
Developer Rationale: We envision several important benefits related to quality improvement with the implementation of this measure. Specifically, the measure will help providers to identify patients who are not adherent (at a critical threshold of 0.8 or greater) to treatment with antipsychotic medications. Guidelines from the American Psychiatric Association (APA) and the National Institute for Clinical Excellence (NICE) emphasize the importance of treatment adherence and uninterrupted antipsychotic regimens to prevent symptoms and relapse. Furthermore, this measure will encourage providers to develop interventions to improve adherence for this high-risk population. The APA guidelines recommend the reasons for nonadherence be considered in the patient's treatment plan. Improved medication adherence would be expected to result in improved symptom control for individuals and a reduction in hospitalizations. Such changes have the potential to improve the quality of care for individuals with schizophrenia and, therefore, advance the quality of care in the area of mental health, a priority area identified by the National Priorities Partnership.

Numerator Statement: Individuals with schizophrenia or schizoaffective disorder who had at least two prescription drug claims for antipsychotic medications and have a PDC of at least 0.8 for antipsychotic medications.
Denominator Statement: Individuals at least 18 years of age as of the beginning of the measurement period with schizophrenia or schizoaffective disorder and at least two prescription drug claims for antipsychotic medications during the measurement period (12 consecutive months).
Denominator Exclusions: Individuals with any diagnosis of dementia during the measurement period.

Measure Type: Process
Data Source: Claims
Level of Analysis: Clinician : Group/Practice, Health Plan, Population : Regional and State

Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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**

Summary of prior review in 2012

- For previous review, the developer provided 21 citations for evidence of high impact in support of the measure.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provides a logic model outlining the process of identifying patients with schizophrenia who are not adherent to antipsychotic medication treatment and the relationship to improved symptom control for those patients identified and a reduction in hospitalization.
- Updated evidence for the measure includes clinical practice guidelines:
  - National Institute for Clinical Excellence (2014) Psychosis and Schizophrenia in Adults: The NICE Guideline on Treatment and Management**, the Guidelines did not provide independent grades to each recommendation**
  - American Psychiatric Association (2010) Practice Guidelines for the Treatment of Patients With Schizophrenia Second Edition, **Overall grades assigned to recommendation were [I] Recommended with substantial clinical confidence and [II] Recommended with moderate clinical confidence.**

Questions for the Committee:

- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

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 performance data from Physician Compare 2015 Individual EP Public Reporting demonstrating some opportunity for improvement.

Year	N	Mean	St Dev	10th	25th	50th	75th	90th	Interquartile Range
2015	80	72.7%	36.4%	10%	33.75%	100%	100%	100%	66.25%

- In addition, the developer provides six studies (Lefeuille et al., 2016; Beebe et al., 2016; Lang et al., 2010; Martin et al., 2009; Ward et al. 2006; Gilmer et al., 2004) that demonstrate low rates of adherence among individuals with schizophrenia who are prescribed antipsychotic medications.

Disparities

- 2007 – 2008 claims data for 36,307 Medicare beneficiaries with schizophrenia were analyzed for disparities.
  - Adherence rates for African-American and Hispanic persons (63.6% and 66.0%) with schizophrenia were substantially lower compared to Whites (79%).
  - Age-related disparities in adherence rates were lower among persons 18 – 44 compared to those over 45.

Questions for the Committee:

- *Do the updated performance data demonstrate 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:

**Updated evidence provided.

**The evidence of % of individuals being prescribed 2 or more antipsychotic medications directly relates to the goal of improving treatment adherence. There is evidence in the literature of improved adherence and better outcomes with long acting injectable antipsychotics as compared to oral medications.

**The evidence is applied indirectly. The submission does not include any information that the measures is improving outcomes or that it is not creating harms.

1b. Performance Gap

Comments:

**Gap continues to exist.

**Yes -- 6 studies demonstrating major problems with adherence to treatment in schizophrenia.

**The measures had a median performance of 100% which indicates that there is no performance gap.

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.

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:

- *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.)?*
- *Is the Committee satisfied with the developers empirical validity testing plan and timeline?*

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

Committee pre-evaluation comments

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

2a1. Reliability – Specifications

Comments:

**There is no reason that this measure cannot be consistently implemented.

**The reliability among physician groups was only adequate if the sample size was greater than 45 patients.

2a2. Reliability – Testing

Comments:

**No concerns.

**No

2b1. Validity –Testing

2b4-7. Threats to Validity

2b4. Meaningful Differences

Comments:

**No concerns.

**As a process measure, this is limited to prescription claims and does not measure whether or not individual patients actually took their medications on a consistent basis -- that is, either missing doses entirely or taking the medication at different times each day.

**Antipsychotics are no indicated for schizoaffective disorder with depression or bipolar. These diagnoses should be removed. Also, consider removing schizophrenia with residual negative symptoms.

2b2-3. Other Threats to Validity

2b2. Exclusions

2b3. Risk Adjustment

Comments:

**Not sure.

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.

- Measure is coded by someone other than person obtaining original information
- All data elements are in defined fields in electronic claims
- Eligible professionals successfully reported this measure to CMS as part of the Physician Quality Reporting Program

Questions for the Committee:

o *Does the Committee have any concerns in regards to the feasibility of the measure?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**Electronic claims; no concerns.

**Use of pharmacy claims data may not capture whether medications are taken by individual patients on a consistent basis.

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**

Accountability program details

- The developer provides three programs that the measure is currently used in:
  - Quality Payment Program (QPP) for which performance results are published on Physician Compare
  - New York State Delivery System Reform Incentive Payment (DSRIP) Program
  - Substance Abuse and Mental Health Services Administration (SAMHSA) Section 223 Demonstration Program

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

- The developer provides a summary of mechanisms for obtaining feedback and feedback from the following programs:
    - Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): No feedback was received specific to this measure.
    - New York State DSRIP Program: No feedback specific to this measure is currently available.
    - SAMHSA Section 223 Demonstration Program: No feedback specific to this measure is currently available.

Additional Feedback:

- The measure went through a re-evaluation process for which feedback from NCQA's measure advisory panels was provided. The panels recommended adding medications which are FDA approved for the treatment of schizophrenia and removing medications which are not FDA approved.

***Questions for the Committee**:*

- *How have 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

- Data from QPP (previously PQRS) was not available at the time of maintenance endorsement to evaluate improvement. Developer plans to provide in future endorsement maintenance reviews.

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

- No unintended consequences were identified during testing, or have been brought to the developers attention since implementation.

Potential harms

- No unintended consequences were identified during testing, or have been brought to the developers attention since implementation.

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**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

**No concerns.

**Not clear. According to the submission, the measure is used in the CMS Quality Payment Program (QPP) and New York DSRIP system (though not publicly reported in the NY program).

**No information on feedback was provided.

4b1. Usability – Improvement

Comments:

**No concerns.

**None stated in the submission. Wide use of this measure has enormous potential to get health plans to better engage in efforts to improve treatment and adherence. Of particular importance would be identifying individual patients that struggle with treatment adherence with oral medications and direct their prescribers to consider long acting injectable (LAI) alternatives.

**No information on unintentional harms appears to have been actively collected. There is no way to judge if the measures is creating harms. Antipsychotics are associated with a number of serious side effects.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0541 : Proportion of Days Covered (PDC): 3 Rates by Therapeutic Category
- 0542 : Adherence to Chronic Medications
- 0543 : Adherence to Statin Therapy for Individuals with Cardiovascular Disease
- 0544 : Use and Adherence to Antipsychotics among members with Schizophrenia
- 0545 : Adherence to Statins for Individuals with Diabetes Mellitus
- 0569 : ADHERENCE TO STATINS
- 1880 : Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder
- Adherence to Antipsychotic Medications for Individuals with Schizophrenia. NCQA is measure steward.

Harmonization

- The developer states that the measure specifications are harmonized with the related measures where possible: proportion of days covered is calculated the same; methodology used to identify denominator; and medications specific to the clinical condition targeted are the same.
- Adherence to Antipsychotic Medications for Individuals with Schizophrenia (NCQA) is used for HEDIS reporting and is harmonized with #1879 in condition, target population, methodology, and medications. The HEDIS measure is only used in Medicaid health plans and therefore is restricted to adults age 18-64.
- 0544 : Use and Adherence to Antipsychotics among members with Schizophrenia is no longer an NQF endorsed measure. Key differences in measure validity and efficiency are addressed in submission.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 1879

Measure Title: Adherence to Antipsychotic Medications for Individuals with Schizophrenia

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document

   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?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐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, skip Questions #3-8, then go to Question #9)

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

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

   Signal-to-noise ratio used to assess variation between state scores

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #6)

   ☒Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

   State level reliability score range .927 - .991; Physician Group (by case volume) mean reliability range .44 - .95

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

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☒Yes (go to Question #7)

   ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2
***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 #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

Reliability at the health plan level was assessed using Cohen's Kappa – measure scores for five randomly selected Medicare Part D plans were compared and inter-rater agreement was calculated.

8. **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?

REFERENCE: Testing attachment, section 2a2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

Health Plan reliability range of Kappa .93 - .97

9. Was **empirical <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

REFERENCE: testing attachment section 2b1.
NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
TIP:** You should answer this question <u>ONLY</u> if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from <u>data element validity testing</u> – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) and <u>all</u> testing results:

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

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

☐Low (please explain below) [NOTE: Should rate <u>LOW</u> 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 <u>not</u> required, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

***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 #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

N/A Process measure

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)

☐Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☒No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with computed performance measure scores for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☐Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***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 #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☐Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☐Yes (go to Question #22)
☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

REFERENCE: *Testing attachment, section 2b1.*

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 #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

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 #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

TIPS*: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1879
Measure Title: Adherence to Antipsychotic Medications for Individuals with Schizophrenia
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: 4/2/2018

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

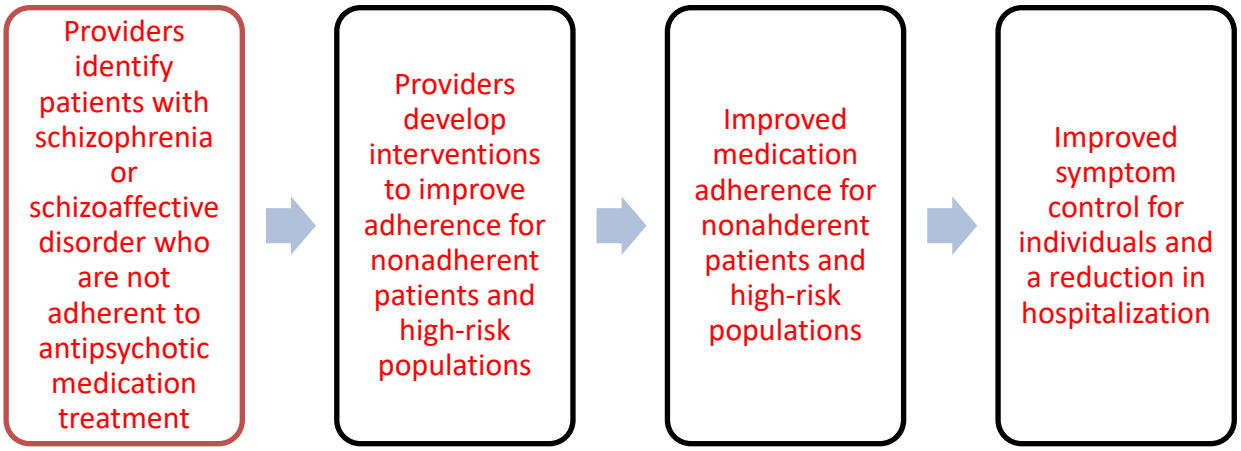
☒ Process: Click here to name what is being measured

☐ 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.



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. This is not a patient-reported measure.

****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)

X 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*)

X Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	National Institute for Clinical Excellence- Psychosis and Schizophrenia in Adults: The NICE Guideline on Treatment and Management National Collaborating Centre for Mental Health 2014 The National Institute for Clinical Excellence and the National Collaborating Centre for Mental health. Psychosis and Schizophrenia in Adults: Prevention and Management. Pages 301-379. Retrieved from https://www.nice.org.uk/guidance/cg178/evidence/full-guideline-pdf-490503565
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.	For people with first episode psychosis offer: • oral antipsychotic medication in conjunction with psychological interventions (family intervention and individual cognitive behavioral therapy). For people with an acute exacerbation or recurrence of psychosis or schizophrenia, offer: • oral antipsychotic medication in conjunction with psychological interventions (family intervention and individual cognitive behavioral therapy). Consider offering depot /long-acting injectable antipsychotic medication to people with psychosis or schizophrenia: • who would prefer such treatment after an acute episode. • where avoiding covert non-adherence (either intentional or unintentional) to antipsychotic medication is a clinical priority within the treatment plan.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline developers used the GRADE system but did not provide independent grades for each recommendation's evidence. The recommendations rely on randomized control trials and meta-analyses, suggesting a high level of quality.
Provide all other grades and definitions from the evidence grading system	Randomized control trials (RCT) without important limitations provide high quality evidence. Observational studies without special strengths or important limitations provide low quality evidence.

	For each outcome, quality may be reduced depending on five factors: methodological limitations, inconsistency, indirectness, imprecision and publication bias.
Grade assigned to the **recommendation** with definition of the grade	The Guidelines did not provide independent grades to each recommendation.
Provide all other grades and definitions from the recommendation grading system	The Guidelines did not provide independent grades to each recommendation.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	For the review of initial treatment with antipsychotic medication: 9 RCTs. For the review of treatment with antipsychotics in people with an acute exacerbation of recurrence of schizophrenia: 72 RCTs. For the review of depot/long-acting injectable antipsychotics: meta-review of five Cochrane reviews.
Estimates of benefit and consistency across studies	There is well-established evidence for the efficacy of antipsychotics in both the treatment of acute psychotic episodes and relapse prevention over time.
What harms were identified?	Side effects of antipsychotics identified include lethargy, sedation, weight gain, sexual dysfunction, and movement disorders.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Not Applicable

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Practice Guidelines for the Treatment of Patients With Schizophrenia Second Edition American Psychiatric Association 2010 Lehman, A. F., Lieberman, J. A., Dixon, L. B., McGlashan, T. H., Miller, A. L., Perkins, and D. O. Kreyenbuhl, J. (2004). Practice Guidelines for the Treatment of Patients with Schizophrenia. American Psychiatric Association. Reprieved from 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.	It is recommended that pharmacological treatment be initiated promptly, provided it will not interfere with diagnostic assessment, because acute psychotic exacerbations are associated with emotional distress, disruption to the patient's life, and a substantial risk of dangerous behaviors to self, others, or property [Recommendation Grade - I].

	While most patients prefer oral medication, patients with recurrent relapses related to nonadherence are candidates for a long-acting injectable antipsychotic medication, as are patients who prefer this mode of administration [Recommendation Grade - II]. If the patient is not improving, it may be helpful to establish whether the lack of response can be explained by medication nonadherence, rapid medication metabolism, or poor absorption [Recommendation Grade - II].
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The attributing evidence is not clearly linked to each recommendation. Each rating of clinical confidence considers the strength of the available evidence and is based on the best available data. When evidence is limited, the level of confidence also incorporates clinical consensus with regard to a particular clinical decision.
Provide all other grades and definitions from the evidence grading system	The following coding system is used to indicate the nature of the supporting evidence in the summary recommendations and references: [A] Double-blind, randomized clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; both the subjects and the investigators are blind to the assignments. [A–] Randomized clinical trial. Same as above but not double-blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally; study does not meet standards for a randomized clinical trial. [C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Case-control study. A study in which a group of patients is identified in the present and information about them is pursued retrospectively or backward in time.
Grade assigned to the **recommendation** with definition of the grade	See brackets after each recommendation above for specific recommendation grades. Overall the grades were: [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence.
Provide all other grades and definitions from the recommendation grading system	The other grade in the recommendation grading system is: [III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	1,272 clinical trials and meta-analyses were screened by using title and abstract information. The Cochrane Database of Systematic Reviews was also searched by using the keyword schizophrenia. Additional, less formal literature searches were conducted by APA staff and individual members of the work group on schizophrenia.

Estimates of benefit and consistency across studies	Nearly all studies found that the antipsychotic medication was superior for treating schizophrenia compared to placebo. These studies demonstrated the efficacy of antipsychotic medications for every subtype and subgroup of patients with schizophrenia. Effectiveness of specific medications will vary by patient symptoms and history.
What harms were identified?	There are numerous side effects to use of both first-generation and second-generation antipsychotics. Antipsychotics are associated with extrapyramidal effects, sedation, orthostatic hypotension and tachycardia, anticholinergic and antiadrenergic effects. Other side effects include weight gain and metabolic effects, and sexual side effects.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No

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.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 1879
Corresponding Measures:
De.2. **Measure Title:** Adherence to Antipsychotic Medications for Individuals with Schizophrenia
Co.1.1. **Measure Steward:** Centers for Medicare and Medicaid Services
De.3. **Brief Description of Measure:** Percentage of individuals at least 18 years of age as of the beginning of the measurement period with schizophrenia or schizoaffective disorder who had at least two prescription drug claims for antipsychotic medications and had a Proportion of Days Covered (PDC) of at least 0.8 for antipsychotic medications during the measurement period (12 consecutive months).
1b.1. **Developer Rationale:** We envision several important benefits related to quality improvement with the implementation of this measure. Specifically, the measure will help providers to identify patients who are not adherent (at a critical threshold of 0.8 or greater) to treatment with antipsychotic medications. Guidelines from the American Psychiatric Association (APA) and the National Institute for Clinical Excellence (NICE) emphasize the importance of treatment adherence and uninterrupted antipsychotic regimens to prevent symptoms and relapse. Furthermore, this measure will encourage providers to develop interventions to improve adherence for this high-risk population. The APA guidelines recommend the reasons for nonadherence be considered in the patient's treatment plan. Improved medication adherence would be expected to result in improved symptom control for individuals and a reduction in hospitalizations. Such changes have the potential to improve the quality of care for individuals with schizophrenia and, therefore, advance the quality of care in the area of mental health, a priority area identified by the National Priorities Partnership.

S.4. **Numerator Statement:** Individuals with schizophrenia or schizoaffective disorder who had at least two prescription drug claims for antipsychotic medications and have a PDC of at least 0.8 for antipsychotic medications.
S.6. **Denominator Statement:** Individuals at least 18 years of age as of the beginning of the measurement period with schizophrenia or schizoaffective disorder and at least two prescription drug claims for antipsychotic medications during the measurement period (12 consecutive months).
S.8. **Denominator Exclusions:** Individuals with any diagnosis of dementia during the measurement period.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Clinician : Group/Practice, Health Plan, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not Applicable. This measure is not paired.

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall

less-than-optimal performance. ***Measures must be judged to meet all 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
1879_Adherence_to_Antipsychotic_Medications_Evidence.docx
1a.1 <u>For Maintenance of Endorsement:</u> 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.
Yes

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)*
<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
We envision several important benefits related to quality improvement with the implementation of this measure. Specifically, the measure will help providers to identify patients who are not adherent (at a critical threshold of 0.8 or greater) to treatment with antipsychotic medications. Guidelines from the American Psychiatric Association (APA) and the National Institute for Clinical Excellence (NICE) emphasize the importance of treatment adherence and uninterrupted antipsychotic regimens to prevent symptoms and relapse. Furthermore, this measure will encourage providers to develop interventions to improve adherence for this high-risk population. The APA guidelines recommend the reasons for nonadherence be considered in the patient's treatment plan. Improved medication adherence would be expected to result in improved symptom control for individuals and a reduction in hospitalizations. Such changes have the potential to improve the quality of care for individuals with schizophrenia and, therefore, advance the quality of care in the area of mental health, a priority area identified by the National Priorities Partnership.

1b.2. **Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis**. *(<u>This is required for maintenance of endorsement</u>. 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.*
PERFORMANCE BASED ON PHYSICIAN QUALITY REPORTING SYSTEM (PQRS) DATA FOR ELIGIBLE PROFESSIONALS (EP):

The following data are extracted from the Physician Compare 2015 Individual EP Public Reporting – Performance Scores file reflecting the most up to date performance data available for this measure. EP performance data is summarized by mean, standard deviation, minimum EP performance, maximum EP performance and performance at 10th, 25th, 50th, 75th, and 90th percentile.

Adherence to antipsychotic medications for individuals with schizophrenia –
YEAR | N | MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | INTERQUARTILE RANGE
2015 | 80 | 72.7% | 36.4% | 10% | 33.75% | 100% | 100% | 100% | 66.25

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.**
OVERVIEW
Six studies (Lefeuille et al., 2016; Beebe et al., 2016; Lang et al., 2010; Martin et al., 2009; Ward et al. 2006; Gilmer et al., 2004) demonstrate low rates of adherence among individuals with schizophrenia who are prescribed antipsychotic medications. These low adherence rates were corroborated by the results of measure testing conducted by FMQAI (now HSAG) of Medicare data, which also showed considerable variation among providers. Both the low rates of adherence and variation among providers indicate a performance gap in the treatment of individuals with schizophrenia. Reported rates of adherence to antipsychotic medications (defined as a PDC or MPR of 0.8 or greater) among persons with schizophrenia range from 41 to 70 percent in these six studies. Martin et al. (2009) suggests that PDC is the most appropriate metric for measuring adherence to antipsychotics. The published studies and the testing results are described below.

PUBLISHED STUDIES:

LAFEUILLE ET AL. (2016): A retrospective study of Medicaid claims between 2008 and 2011 from 5 states found that among nearly 13,000 patients who received antipsychotics during the study period, 48.6 percent met the HEDIS measure's (Adherence to Antipsychotic Medications for Individuals with Schizophrenia) criteria for achieved continuity (PDC =80 percent). Rates were similar between patients receiving paliperidone palmitate (46.3 percent) and those receiving other antipsychotics (48.7 percent). Patients that met continuity criteria during the baseline year were more likely to be adherent in the measurement year (76.2 percent) than patients non-adherent in the baseline year (27.3 percent) ($p<0.001$).

BEEBE ET AL. (2016): One cross sectional descriptive study on 185 stable outpatients (i.e. did not include first episode participants) with schizophrenia spectrum disorders found adherence to antipsychotics determined through pill counts ranged from 0 to 100 percent with a mean of 70 percent (SD 34.9).

LANG ET AL. (2010): A recent study (Lang et al., 2010) was a retrospective analysis using claims data (July 1, 2004 - June 30, 2005) that identified 12,032 Florida Medicaid recipients with a diagnosis of schizophrenia who were prescribed an antipsychotic medication and were followed for one year after the prescription. During the one-year follow-up, only 66 percent of patients were adherent (MPR 80 percent or greater), 20 percent were partially adherent (MPR greater than or equal to 50 percent and less than 80 percent), and 14 percent were non-adherent (MPR < 50 percent).

MARTIN ET AL. (2009): Using data for patients with schizophrenia, this retrospective study analyzed North Carolina Medicaid administrative claims data from July 1999 to June 2000 with a final sample of 25,200 person-quarters with data from 7069 individuals. The study demonstrated that PDC was a more conservative metric compared to MPR and recommended that for drug classes such as antipsychotics the PDC should be used to measure adherence. The result of the analysis for PDC of patients that were adherent (PDC of 0.8 or greater) by quarter was approximately 41 percent.

WARD ET AL. (2006): A third study (Ward et al., 2006) was also a retrospective analysis of persons diagnosed with schizophrenia in two Canadian provinces. The level of adherence to the atypical antipsychotic medications (risperidone, olanzapine, or quetiapine) was measured among 41,754 and 3,291 patients in Quebec and Saskatchewan, respectively. During the follow-up period (mean of 2.6 and 3.1 years in Quebec and Saskatchewan, respectively), only 61.4 percent (Quebec) and 45.1 percent (Saskatchewan) of patients had good adherence (MPR 80 percent or greater).

GILMER ET AL. (2004): Similarly, a fourth study (Gilmer et al., 2004) was a retrospective study that analyzed adherence to antipsychotic medications for persons with schizophrenia in San Diego County, representing 2801 person-years. Using Medicaid claims data for fiscal years 1999 and 2000, they found that only 41 percent of patients were adherent (MPR 80 percent or greater), 16 percent were partially adherent (MPR greater than or equal to 50 percent and less than 80 percent), and 24 percent were non-adherent (MPR < 50 percent) during the year following study enrollment.

References:

Beebe, L. H., Smith, K., and Phillips, C. (2016) Descriptions and correlates of medication adherence, attitudes, and self-efficacy in outpatients with schizophrenia spectrum disorders (SSDs). Archives of Psychiatric Nursing, 30(3), 400-405.

Gilmer, T. P., Dolder, C. R., Lacro, J. P., Folsom, D. P., Lindamer, L., Garcia, P., et al. (2004). Adherence to treatment with antipsychotic medication and health care costs among Medicaid beneficiaries with schizophrenia. American Journal of Psychiatry, 161(4), 692-9.

Lafeuille, M., Frois, C., Cloutier, M., Duh, M. S., Lefebvre, P., Pesa, J., and ... Durkin, M. (2016). Factors associated with adherence to the HEDIS quality measure in Medicaid patients with schizophrenia. American Health and Drug Benefits, 9(7), 399-409.

Lang, K., Meyers, J. L., Korn, J. R., Lee, S., Sikirica, M., Crivera, C., et al. (2010). Medication adherence and hospitalization among patients with schizophrenia treated with antipsychotics. Psychiatr Serv, 61(12), 1239-1247.

Martin, B. C., Wiley-Exley, E. K., Richards, S., Domino, M. E., Carey, T. S., and Sleath, B. L. (Jan 2009). Contrasting measures of adherence with simple drug use, medication switching, and therapeutic duplication. Ann Pharmacother, 43(1), 36-44.

Ward, A., Ishak, K., Proskorovsky, I., and Caro, J. (2006). Compliance with refilling prescriptions for atypical antipsychotic agents and its association with the risks for hospitalization, suicide, and death in patients with schizophrenia in Quebec and Saskatchewan: A retrospective database study. Clin Ther, 28(11), 1912-1921.

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.*

We analyzed 2007-2008 claims data for 36,307 Medicare beneficiaries with schizophrenia. A consistent pattern was observed with adherence rates for antipsychotic medication being substantially lower among African-American and Hispanic persons with schizophrenia compared with Whites. For all age groups combined, the adherence rates were 63.6 percent and 66.0 percent for African-American and Hispanic persons, respectively as compared to, 79.0 percent for White persons. Additionally, adherence rates were lower among African-American and Hispanic persons than among White persons in every age group.

In regard to age-related disparities, adherence rates were lower among persons 18-44 years of age (i.e., 64.8 percent (18-24 years) and 70.8 percent (25-44 years)) as compared to those over 45 years of age (i.e., 77.6 percent (45-64 years), 76.5 percent (65-74 years), 77.8 percent (75-84 years), and 77.8 percent (85 years and older)). This pattern of lower adherence rates in younger persons was generally consistent across ethnic groups (White, African-American, and Hispanic persons).

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

Substantial disparities in adherence rates for antipsychotic medications have been observed between race/ethnicity groups and age groups among persons with schizophrenia in published studies and in our testing results. One recent study did not find significant associations between adherence and patient characteristics in stable outpatients (i.e. patients who are not experiencing their first episode of psychosis).

PUBLISHED STUDIES

Six studies described in this section (Garcia et al., 2016; Lafeuille et al., 2016; Beebe et al., 2016; Busch, Lehman, Goldman, and Frank, 2009; Ahn et al., 2008; Gilmer et al., 2004; Valenstein et al., 2004) reported lower adherence rates among African-American or Hispanic persons with schizophrenia as compared to White persons with Schizophrenia. One study did not find significant differences among racial/ethnic groups in stable outpatients (Beebe et al., 2016).

GARCIA ET AL. (2016): This systematic review found age, race, and education to be associated with adherence rates. Younger patients were less adherent than older patients, black patients had lower adherence rates than white patient, and patients with lower levels of education had poorer adherence. The review found economic and transportation barriers hinder patient's adherence to treatment.

LAFEUILLE ET AL. (2016): A retrospective study of claims between 2008 and 2011 from 5 states found women (OR, 1.11; 95% CI, 1.01-1.22), age 55 to 64 compared to age 25-34 (OR, 1.26; 95% CI, 1.09-1.46), and Hispanic ethnicity compared to White (OR, 1.37; 95% CI,1.05-1.81) were associated with higher odds of meeting continuity criteria (PDC > 0.8) for the Adherence to Antipsychotic Medications for Individuals with Schizophrenia HEDIS measure.

BEEBE ET AL. (2016): One study on 185 stable outpatients (i.e. patients who are not experiencing their first episode of psychosis) with schizophrenia spectrum disorders found no significant associations between adherence and age, diagnosis, gender, race, or education level.

BUSCH ET AL. (2009): In an observational study based on five years of claims data (July 1, 1996 to June 30, 2001), Busch et al. (2009) assessed quality of care related to the treatment of schizophrenia among 23,619 Medicaid enrollees in Florida. In comparing African-American patients with White patients in the maintenance phase, they reported a significantly lower rate among African-Americans for a measure related to adherence (i.e., having a continuous supply of an antipsychotic medication) (odds ratio 0.56; 95% CI 0.53-0.60).

AHN ET AL. (2008): In an analysis of 1994-2003 Medicaid claims data for 36,195 individuals with schizophrenia in California, being classified as non-adherent (defined using a medication possession ratio and other variables) was associated with being African-American or Hispanic.

GILMER ET AL. (2004): In a retrospective study using Medicaid claims data for fiscal years 1999 and 2000 in San Diego County (N=2801 person-years), the rate of adherence (MPR 0.8 or greater) was lower among African-Americans (34.9 percent) than among Whites (42.8 percent) or Hispanics (36.9 percent).

VALENSTEIN ET AL. (2004): In a claims-based study of 49,003 veterans with schizophrenia taking one antipsychotic medication during 12 months in 1998-1999, 54 percent of African-Americans were poorly adherent (MPR less than 0.8) compared to 32 percent of Whites in a descriptive analysis; in a logistic regression analysis, the odds ratio comparing the risk of poor adherence among African-Americans to Whites was 2.38 (95% CI 2.28-2.49).

References:

Ahn, J., McCombs, J. S., Jung, C., Croudace, T. J., McDonnell, D., Ascher-Svanum, H., et al. (2008). Classifying patients by antipsychotic adherence patterns using latent class analysis: Characteristics of nonadherent groups in the California Medicaid (Medi-Cal) Program. Value in Health, 11(1), 48-56.

Beebe, L. H., Smith, K., and Phillips, C. (2016) Descriptions and correlates of medication adherence, attitudes, and self-efficacy in outpatients with schizophrenia spectrum disorders (SSDs). Archives of Psychiatric Nursing, 30(3), 400-405.

Busch, A. B., Lehman, A. F., Goldman, H., and Frank, R. G. (2009). Changes over time and disparities in schizophrenia treatment quality. Medical Care, 47(2), 199-207.

Garcia, S., Martínez-Cengotitabengoa, M., López-Zurbano, S., et al. (2016). Adherence to antipsychotic medication in bipolar disorder and schizophrenic patients: a systematic review. Journal of Clinical Psychopharmacology, 36(4), 355-371.

Gilmer, T. P., Dolder, C. R., Lacro, J. P., Folsom, D. P., Lindamer, L., Garcia, P., et al. (2004). Adherence to treatment with antipsychotic medication and health care costs among Medicaid beneficiaries with schizophrenia. American Journal of Psychiatry, 161(4), 692-9.

Lafeuille, M., Frois, C., Cloutier, M., Duh, M. S., Lefebvre, P., Pesa, J., and ... Durkin, M. (2016). Factors associated with adherence to the HEDIS quality measure in Medicaid patients with schizophrenia. American Health and Drug Benefits, 9(7), 399-409.

Valenstein, M., Blow, F. C., Copeland, L. A., McCarthy, J. F., Zeber, J. E., Gillon, L., et al. (2004). Poor antipsychotic adherence among patients with schizophrenia: Medication and patient factors. Schizophrenia Bulletin, 30(2), 255-64.

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):*
Behavioral Health

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

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Elderly, Populations at Risk, Populations at Risk : Dual eligible beneficiaries

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Measure #383 at https://www.cms.gov/Medicare/Quality-Payment-Program/Resource-Library/2017-Resources.html

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:** NQF_1879_Code_Tables_2018_Final.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.
Yes

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.
- Updated NDCs as of February 20, 2018
- Removed medications lacking FDA approval for treatment of schizophrenia: pimozide and olanzapine-fluoxetine
- Added medications with FDA approval for treatment of schizophrenia: cariprazine, quetiapine fumarate (Seroquel), brexpiprazole, aripiprazole lauroxil (Aristada)

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).
Individuals with schizophrenia or schizoaffective disorder who had at least two prescription drug claims for antipsychotic medications and have a PDC of at least 0.8 for antipsychotic medications.

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 defined as individuals with a PDC of 0.8 or greater.

The PDC is calculated as follows:

PDC NUMERATOR
The PDC numerator is the sum of the days covered by the days' supply of all prescription drug claims for all antipsychotic medications. The period covered by the PDC starts on the day the first prescription is filled (index date) and lasts through the end of the measurement period, or death, whichever comes first. For prescription drug claims with a days' supply that extends beyond the end of the measurement period, count only the days for which the drug was available to the individual during the measurement period. If there are claims for the same drug (generic name) on the same date of service, keep the claim with the

largest days' supply. If claims for the same drug (generic name) overlap, then adjust the prescription start date to be the day after the previous fill has ended.

PDC DENOMINATOR
The PDC denominator is the number of days from the first prescription drug claim date through the end of the measurement period, or death date, whichever comes first.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Individuals at least 18 years of age as of the beginning of the measurement period with schizophrenia or schizoaffective disorder and at least two prescription drug claims for antipsychotic medications during the measurement period (12 consecutive months).

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. Continuously enrolled in Medicare Part D with no more than a one-month gap in enrollment during the measurement period;
2. Continuously enrolled in Medicare Part A and Part B with no more than a one-month gap in Part A enrollment and no more than a one-month gap in Part B enrollment during the measurement period; and,
3. No more than one month of HMO (Health Maintenance Organization) enrollment during the measurement period.

IDENTIFICATION OF SCHIZOPHRENIA
Individuals with schizophrenia or schizoaffective disorder are identified by having a diagnosis of schizophrenia within the inpatient or outpatient claims data. Individuals must have:

At least two encounters with a diagnosis of schizophrenia or schizoaffective disorder with different dates of service in an outpatient setting, emergency department setting, or non-acute inpatient setting during the measurement period;

OR

At least one encounter with a diagnosis of schizophrenia or schizoaffective disorder in an acute inpatient setting during the measurement period.

CODES USED TO IDENTIFY SCHIZOPHRENIA OR SCHIZOAFFECTIVE DISORDER DIAGNOSIS
Codes used to identify schizophrenia or schizoaffective disorder are included in the attached excel worksheet of codes (NQF_1879_Code Tables_2018_Final.xlsx) under the tab NQF_1879_Schizophrenia.

Table 1: Schizophrenia or Schizoaffective Disorder Diagnosis
ICD-9-CM: 295.xx
ICD-10-CM: F20.0, F20.1, F20.2, F20.3, F20.5, F20.81, F20.89, F20.9, F25.0, F25.1, F25.8, F25.9

CODES USED TO IDENTIFY ENCOUNTER TYPE:
Codes used to identify encounters are under tab NQF_1879_Encounter_types.

Table 2.1: Outpatient Setting
Current Procedural Terminology (CPT): 98960-98962, 99078, 99201-99205, 99211-99215, 99217-99220, 99241-99245, 99341-99345, 99347-99350, 99385-99387, 99395-99397, 99401-99404, 99411, 99412, 99429, 99510
HCPCS: G0155, G0176, G0177, G0409-G0411, G0463, H0002, H0004, H0031, H0034-H0037, H0039, H0040, H2000, H2001, H2010-H2020, M0064, S0201, S9480, S9484, S9485, T1015
UB-92 revenue: 0510, 0511, 0513, 0516-0517, 0519-0523, 0526-0529, 0770, 0771, 0779, 0900-0905, 0907, 0911-0917, 0919, 0982, 0983

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 90880, 99221-99223, 99231-99233, 99238, 99239, 99251-99255, 99291

WITH

Place of Service (POS): 03, 05, 07, 09, 11, 12, 13, 14, 15, 20, 22, 24, 33, 49, 50, 52, 53, 71, 72

Table 2.2: Emergency Department Setting
CPT: 99281-99285
UB-92 revenue: 0450, 0451, 0452, 0456, 0459, 0981

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 99291

WITH

POS: 23

Table 2.3: Non-Acute Inpatient Setting
CPT: 99304-99310, 99315, 99316, 99318, 99324-99328, 99334-99337
HCPCS: H0017-H0019, T2048
UB-92 revenue: 0118, 0128, 0138, 0148, 0158, 0190-0194, 0199, 0524, 0525, 0550-0552, 0559, 0660-0663, 0669, 1000, 1001, 1003-1005

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 99291

WITH

POS: 31, 32, 56

Table 2.4: Acute Inpatient Setting
UB-92 revenue: 0100, 0101, 0110-0114, 0119-0124, 0129-0134, 0139-0144, 0149-0154, 0159, 0160, 0164, 0167, 0169, 0200-0204, 0206-0209, 0210-0214, 0219, 0720-0724, 0729, 0987

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 99221-99223, 99231-99233, 99238, 99239, 99251-99255, 99291

WITH

POS: 21, 51

IDENTIFICATION OF PRESCRIPTION DRUG CLAIMS FOR ANTIPSYCHOTIC MEDICATION:

Individuals with at least two prescription drug claims for any of the following oral antipsychotic medications (Table 3: Oral Antipsychotic Medications) or long-acting injectable antipsychotic medications (see Table 4: Long-acting injectable antipsychotic medications). The National Drug Center (NDC) identifier for medications included in the measure denominator are listed in tab NQF_1879_ Antipsychotics of the attached excel workbook. Obsolete drug products are excluded from National Drug Codes (NDCs) with an inactive date more than six years prior to the beginning of the measurement period or look-back period.

TABLE 3: ORAL ANTIPSYCHOTIC MEDICATIONS
The following are oral formulations only.

Typical Antipsychotic Medications:
chlorpromazine
fluphenazine
haloperidol
loxapine
molindone
perphenazine
prochlorperazine
thioridazine
thiothixene
trifluoperazine

Atypical Antipsychotic Medications:
aripiprazole
asenapine
brexpiprazole
cariprazine
clozapine
iloperidone
lurasidone
olanzapine
paliperidone
quetiapine
quetiapine fumarate (Seroquel)
risperidone
ziprasidone

Antipsychotic Combinations:
perphenazine-amitriptyline

TABLE 4: LONG-ACTING INJECTABLE ANTIPSYCHOTIC MEDICATIONS
The following are the long-acting (depot) injectable antipsychotic medications by class for the denominator. The route of administration includes all injectable and intramuscular formulations of the medications listed below.

Typical Antipsychotic Medications:
fluphenazine decanoate (J2680)
haloperidol decanoate (J1631)

Atypical Antipsychotic Medications:
aripiprazole (J0401)
aripiprazole lauroxil (Aristada)
olanzapine pamoate (J2358)
paliperidone palmitate (J2426)
risperidone microspheres (J2794)

Note: Since the days' supply variable is not reliable for long-acting injections in administrative data, the days' supply is imputed as listed below for the long-acting (depot) injectable antipsychotic medications billed under Medicare Part D and Part B:
fluphenazine decanoate (J2680) – 28 days' supply
haloperidol decanoate (J1631) – 28 days' supply
aripiprazole (J0401) – 28 days' supply
aripiprazole lauroxil (Aristada) - 28 days' supply
olanzapine pamoate (J2358) – 28 days' supply
paliperidone palmitate (J2426) – 28 days' supply
risperidone microspheres (J2794) – 14 days' supply

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

Individuals with any diagnosis of dementia during the measurement period.

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.)*
Individuals with any diagnosis of dementia are identified with the diagnosis codes listed below tab NQF_1879_Dementia

Table 5: Codes Used to Identify Dementia
ICD-9-CM: 290.0, 290.10, 290.11, 290.12, 290.13, 290.20, 290.21, 290.3, 290.40, 290.41, 290.42, 290.43, 290.8, 290.9, 291.2, 292.82, 294.10, 294.11, 294.20, 294.21, 330.1, 331.0, 331.19, 331.82
ICD-10-CM: E75.00, E75.01, E75.02, E75.09, E75.10, E75.11, E75.19, E75.4, F01.50, F01.51, F02.80, F02.81, F03.90, F03.91, F05, F10.27, F11.122, F13.27, F13.97, F18.17, F18.27, F18.97, F19.17, F19.27, F19.97, G30.0, G30.1, G30.8, G30.9, G31.09, G31.83

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.)*
Depending on the operational use of the measure, measure results can be stratified by:
- State
- Physician Group*
- Age – Divided into six categories: 18-24, 25-44, 45-64, 65-74, 75-84, and 85+ years
- Race/Ethnicity
- Dual Eligibility

*See Calculation Algorithm/Measure Logic S.14 below for physician group attribution methodology used for this measure.

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.)*
Target Population: Individuals at least 18 years of age as of the beginning of the measurement period who have met the enrollment criteria for Medicare Parts A, B, and D.

Denominator: Individuals at least 18 years of age as of the beginning of the measurement period with schizophrenia or schizoaffective disorder and at least two prescription drug claims for antipsychotic medications during the measurement period (12 consecutive months).

CREATE DENOMINATOR:
1. Pull individuals who are 18 years of age or older as of the beginning of the measurement period.
2. Include individuals who were continuously enrolled in Medicare Part D coverage during the measurement period, with no more than a one-month gap in enrollment during the measurement period, or up until their death date if they died during the measurement period.
3. Include individuals who had no more than a one-month gap in Medicare Part A enrollment, no more than a one-month gap in Part B enrollment, and no more than one month of HMO (Health Maintenance Organization) enrollment during the current measurement period (fee-for-service [FFS] individuals only).
4. Of those individuals identified in Step 3, keep individuals who had:

At least two encounters with a diagnosis of schizophrenia of schizoaffective disorder with different dates of service in an outpatient setting, emergency department setting, or non-acute inpatient setting during the measurement period;
OR
Individuals who had at least one encounter with a diagnosis of schizophrenia or schizoaffective disorder in an acute inpatient setting during the measurement period.

5. For the individuals identified in Step 4, extract Medicare Part D claims for any antipsychotic medication during the measurement period. Attach the generic name and the drug ID to the dataset.
6. Of the individuals identified in Step 5, exclude those who did not have at least two prescription drug claims for any antipsychotic medication on different dates of service (identified by having at least two Medicare Part D claims with the specific codes) during the measurement period.
7. Exclude those individuals with a diagnosis of dementia during the measurement period.

Numerator: Individuals with schizophrenia or schizoaffective disorder who had at least two prescription drug claims for antipsychotic medications and have a PDC of at least 0.8 for antipsychotic medications.

CREATE NUMERATOR:
For the individuals in the denominator, calculate the PDC for each individual according to the following methods:
1. Determine the individual's medication therapy period, defined as the number of days from the index prescription date through the end of the measurement period, or death, whichever comes first. The index date is the service date (fill date) of the first prescription drug claim for an antipsychotic medication in the measurement period.
2. Within the medication therapy period, count the days the individual was covered by at least one drug in the antipsychotic medication class based on the prescription drug claim service date and days of supply.
a. Sort and de-duplicate Medicare Part D antipsychotic medication claims by beneficiary ID, service date, generic name, and descending days' supply. If prescriptions for the same drug (generic name) are dispensed on the same date of service for an individual, keep the dispensing with the largest days' supply.
b. Calculate the number of days covered by antipsychotic drug therapy per individual.
i. For prescription drug claims with a days' supply that extends beyond the end of the measurement period, count only the days for which the drug was available to the individual during the measurement period.
ii. If claims for the same drug (generic name) overlap, then adjust the prescription start date to be the day after the previous fill has ended.
iii. If claims for different drugs (different generic names) overlap, do not adjust the prescription start date.
3. Calculate the PDC for each individual. Divide the number of covered days found in Step 2 by the number of days in the individual's medication therapy period found in Step 1.

An example of SAS code for Steps 1-3 was adapted from Pharmacy Quality Alliance (PQA) and is available at the URL: http://www2.sas.com/proceedings/forum2007/043-2007.pdf.

4. Of the individuals identified in Step 3, count the number of individuals with a calculated PDC of at least 0.8 for the antipsychotic medications. This is the numerator.

PHYSICIAN GROUP ATTRIBUTION:
Physician group attribution was adapted from Generating Medicare Physician Quality Performance Measurement Results (GEM) Project: Physician and Other Provider Grouping and Patient Attribution Methodologies (http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/GEM/downloads/GEMMethodologies.pdf). The following is intended as guidance and reflects only one of many methodologies for assigning individuals to a medical group. Please note that the physician group attribution methodology excludes patients who died, even though the overall measure does not.

I. Identify Physician and Medical Groups
1. Identify all Tax Identification Numbers (TINs)/National Provider Identification (NPIs) combinations from all Medicare Part B claims in the measurement year and the prior year. Keep records with valid NPI. Valid NPIs have 10 numeric characters (no alpha characters).
2. For valid NPIs, pull credentials and specialty code(s) from the CMS provider tables.
3. Create one record per NPI with all credentials and all specialties. A provider may have more than one specialty.
4. Attach TIN to NPI, keeping only those records with credentials indicating a physician (MD or DO), physician assistant (PA), or nurse practitioner (NP).

5. Identify medical group TINs: Medical group TINs are defined as TINs that had physician, physician assistant, or nurse practitioner provider specialty codes on at least 50% of Medicare Part B carrier claim line items billed by the TIN during the measurement year or prior year. (The provider specialty codes are listed after Patient Attribution.)
a. Pull Part B records billed by TINS identified in Step 4 during the measurement year and prior year.
b. Identify claims that had the performing NPI (npi_prfrmg) in the list of eligible physicians/TINs, keeping those that match by TIN, performing NPI, and provider state code.
c. Calculate the percentage of Part B claims that match by TIN, npi_prfrmg, and provider state code for each TIN, keeping those TINs with percentages greater than or equal to 50%.
d. Delete invalid TINs. Examples of invalid TINs are defined as having the same value for all nine digits or values of 012345678, 012345678, 123456789, 987654321, or 87654321.
6. Identify TINs that are not solo practices.
a. Pull Part B records billed by physicians identified in Step 4 for the measurement year and/or prior year.
b. Count unique NPIs per TIN.
c. Keep only those TINs having two or more providers.
d. Delete invalid TINs. Examples of invalid TINs are defined as having the same value for all nine digits or values of 012345678, 012345678, 123456789, 987654321, or 87654321.
7. Create final group of TINs from Step 5 and Step 6 (TINs that are medical groups and are not solo practices).
8. Create file of TINs and NPIs associated with those TINs. These are now referred to as the medical group TINs.
9. Determine the specialty of the medical group (TIN) to be used in determining the specialty of nurse practitioners and physician assistants. The plurality of physician providers in the medical group determines the specialty of care for nurse practitioners and physician assistants.
a. From the TIN/NPI list created in Step 8, count the NPIs per TIN/specialty.
b. The specialty with the maximum count is assigned to the medical group.

II. Identify Individual Sample and Claims
10. Create individual sample.
a. Pull individuals with 11+ months of Medicare Parts A, B, and D during the measurement year.
b. Verify the individual did not have any months with Medicare as secondary payer. Remove individuals with BENE_PRMRY_PYR_CD not equal to one of the following:
• A = working-age individual/spouse with an employer group health plan (EGHP)
• B = End Stage Renal Disease (ESRD) in the 18-month coordination period with an EGHP
• G = working disabled for any month of the year
c. Verify the individual resides in the U.S., Puerto Rico, Virgin Islands, or Washington D.C.
d. Exclude individuals who enter the Medicare hospice at any point during the measurement year.
e. Exclude individuals who died during the measurement year.
11. For individuals identified in Step 10, pull office visit claims that occurred during the measurement year and in the six months prior to the measurement year.
a. Office visit claims have CPT codes of 99201-99205, 99211-99215, and 99241-99245.
b. Exclude claims with no npi_prfrmg.
12. Attach medical group TIN to claims by NPI.

III. Patient Attribution
13. Pull all Medicare Part B office claims from Step 12 with specialties indicating primary care or psychiatry (see list of provider specialties and specialty codes below). Attribute each individual to at most one medical group TIN for each measure.
a. Evaluate specialty on claim (HSE_B_HCFA_PRVDR_SPCLTY_CD) first. If specialty on claim does not match any of the measure-specific specialties, then check additional specialty fields.
b. If the provider specialty indicates nurse practitioners or physician assistants (code 50 or code 97), then assign the medical group specialty determined in Step 9.
14. For each individual, count claims per medical group TIN. Keep only individuals with two or more E&M claims.
15. Attribute individual to the medical group TIN with the most claims. If a tie occurs between medical group TINs, attribute the TIN with the most recent claim.
16. Attach the medical group TIN to the denominator and numerator files by individual.

Provider Specialties and Specialty Codes
Provider specialties and specialty codes include only physicians, physician assistants, and nurse practitioners for physician grouping, TIN selection, and patient attribution. The provider specialty codes and the associated provider specialty are shown below:

01—General practice*
02—General surgery
03—Allergy/immunology
04—Otolaryngology
05—Anesthesiology
06—Cardiology
07—Dermatology
08—Family practice*
09—Interventional pain management
10—Gastroenterology
11—Internal medicine*
12—Osteopathic manipulative therapy
13—Neurology
14—Neurosurgery
16—Obstetrics/gynecology*
18—Ophthalmology
20—Orthopedic surgery
22—Pathology
24—Plastic and reconstructive surgery
25—Physical medicine and rehabilitation
26—Psychiatry*
28—Colorectal surgery
29—Pulmonary disease
30—Diagnostic radiology
33—Thoracic surgery
34—Urology
37—Nuclear medicine
38—Geriatric medicine*
39—Nephrology
39—Pediatric medicine
40—Hand surgery
44—Infectious disease
46—Endocrinology
50—Nurse practitioner*
66—Rheumatology
70—Multi-specialty clinic or group practice*
72—Pain management
76—Peripheral vascular disease
77—Vascular surgery
78—Cardiac surgery
79—Addiction medicine
81—Critical care (intensivists)
82—Hematology
83—Hematology/oncology
84—Preventive medicine*
85—Maxillofacial surgery
86—Neuropsychiatry*
90—Medical oncology
91—Surgical oncology
92—Radiation oncology
93—Emergency medicine
94—Interventional radiology
97—Physician assistant*
98—Gynecologist/oncologist
99—Unknown physician specialty
Other—NA

*Provider specialty codes specific to this measure

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 does not use a sample or survey.

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.

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.
The data source for the measure calculation required the following Medicare files depending on the level of accountability where the measure is being used:
• Denominator tables to determine individual enrollment
• Prescription drug benefit (Part D) coverage tables
• Beneficiary file
• Institutional claims (Part A)
• Non-institutional claims (Part B)—physician carrier/non-DME (durable medical equipment)
• Prescription drug benefit (Part D) claims
• Centers for Medicare and Medicaid Services (CMS) physician and physician specialty tables
• National Plan and Provider Enumeration System (NPPES) database

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, Health Plan, 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.)*

2. Validity – See attached Measure Testing Submission Form
1879_Adherence_to_Antipsychotic_Medications_Testing.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.
Yes

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.
No

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 1879
Measure Title: Adherence to Antipsychotic Medications for Individuals with Schizophrenia
Date of Submission: 4/2/2018
Type of Measure:

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

Instructions

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

Medicare Parts A, B, and D claims data and Minimum Data Set (MDS) data were used to support the field testing of the measure. The following files were used:

- Denominator tables to determine individual enrollment
- Prescription drug benefit (Part D) coverage tables
- Beneficiary file
- Institutional claims (Part A)
- Non-institutional claims (Part B)—physician carrier/non-DME (durable medical equipment)
- Prescription drug benefit (Part D) claims
- Centers for Medicare & Medicaid Services (CMS) physician and physician specialty tables
- National Plan & Provider Enumeration System (NPPES) database

1.3. What are the dates of the data used in testing? 2007, 2008

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: population (state)	☒ other: population (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*)

Data from eight states were included in the testing and analysis for validity and physician group and state reliability. These data included 9,406 Physician Groups and 656 Part D plans.

For health plan reliability testing, data included five randomly selected Part D plans from two states.

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*)

The data included 4,789,034 Medicare beneficiaries.

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.

No differences in the data or sample used.

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.

Two proxy variables for social risk were evaluated to understand disparities: race/ethnicity and dual-eligibility beneficiary status. Because this measure is not an outcome or intermediate outcome measure, these factors were not evaluated for risk adjustment. Overall, in the younger age groups (18-64), African-Americans had noticeably lower adherence. In all age groups, dual-eligible beneficiaries had higher rates of adherence than those who are not dual-eligible.

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*)

In order to assess measure precision in the context of the observed variability across measurement units (physician groups), we utilized the approach proposed by Adams (2009) in work on the reliability for provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient."

The signal to noise ratio was calculated as a function of the variance between physician groups (signal) and the variance within a physician group (noise). Reliability was estimated using a beta-binomial model. This approach has 2 basic assumptions:

1) Each physician has a true pass rate, p, which varies from physician to physician, and
2) The physician's score is a binomial random variable conditional on the physician's true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual physician group variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across physician groups). In a simulation, Adams showed that differences between physicians started to be seen at reliability of 0.7 and significant differences could be seen at reliability of 0.9. Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between physicians. Reliability scores were also calculated for state level results using the same approach.

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

Reliability at the health plan level was assessed using Cohen's Kappa. The measure scores for five randomly selected Medicare Part D plans were compared and inter-rater agreement was calculated. Concerning an acceptable threshold for kappa, there are no definitive criteria in the literature for what level of reliability is acceptable for measures based on administrative data. Furthermore, since relatively small differences in programmer interpretation could result in a large variation in output, we utilized a conservative threshold of 0.9 for Cohen's Kappa, based on the following scale:

< 0 = no agreement
0–0.20 = slight agreement
0.21–0.40 = fair agreement
0.41–0.60 = moderate agreement
0.61–0.80 = substantial agreement
0.81–1 = almost perfect agreement

Therefore, if the Cohen's Kappa was greater than or equal to 0.9, the measure specifications were considered reliable. If Cohen's Kappa in the initial reliability testing with the two programmers was less than 0.9, each step of the measure algorithm (in the Measure Information Form [MIF]) was compared, and the differences were clarified between programmer 1 and 2. Identified differences are noted in a narrative, where applicable, along with extracts of the respective modification to the MIF.

The revised MIF was then presented to a third programmer and results compared to the consolidated results derived in the first round of reliability testing. This iterative process with independent programmers continued until the Kappa score reached the threshold of greater than or equal to 0.9.

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*)

State Reliability

State / Denominator / Mean rate for state / Reliability score (based on the mean rate)

A / 1368 / 67.54% / 0.955

B / 681 / 76.36% / 0.927

C / 14869 / 71.03% / 0.996

D / 3652 / 84.72% / 0.990

E / 6157 / 80.02% / 0.992

F / 3351 / 68.49% / 0.981

G / 1005 / 78.31% / 0.952

H / 5224 / 81.13% / 0.991

Physician Group Reliability (By Case Volume)

Minimum denominator size of MD group / # of Groups / Mean rate of physician groups / Variance between physician groups / Physician specific error / Reliability score (based on the mean rate and the minimum denominator size) / Mean Reliability Score / Median Reliability score / Minimum Reliability Score / Maximum Reliability Score / Standard Deviation of Reliability Scores

10 / 296 / 76.71% / 0.0081 / 0.0179 / 0.3116 / 0.48 / 0.44 / 0.26 / 0.91 / 0.15

20 / 122 / 77.49% / 0.0087 / 0.0087 / 0.4993 / 0.65 / 0.63 / 0.43 / 0.91 / 0.13

30 / 71 / 79.08% / 0.0079 / 0.0055 / 0.5895 / 0.71 / 0.71 / 0.52 / 0.91 / 0.11

35 / 55 / 80.28% / 0.0081 / 0.0045 / 0.6405 / 0.75 / 0.74 / 0.58 / 0.91 / 0.09

40 / 44 / 80.94% / 0.0088 / 0.0039 / 0.6954 / 0.79 / 0.8 / 0.64 / 0.92 / 0.08

45 / 36 / 81.41% / 0.0084 / 0.0034 / 0.7144 / 0.8 / 0.8 / 0.67 / 0.92 / 0.07

50 / 30 / 80.68% / 0.0092 / 0.0031 / 0.7471 / 0.82 / 0.83 / 0.69 / 0.92 / 0.06

100 / 7 / 74.55% / 0.0194 / 0.0019 / 0.9107 / 0.94 / 0.95 / 0.91 / 0.96 / 0.02

150 / 3 / 75.47% / 0.0032 / 0.0012 / 0.7186 / 0.75 / 0.76 / 0.71 / 0.77 / 0.03

Health Plan Reliability

	Percent Agreement		
Unit of Analysis	**Programmer 1 Num/Den (%)**	**Programmer 2 Num/Den (%)**	**Final Cohen's Kappa**
Part D Plan 1	44/75 (58.7%)	45/75 (60.0%)	0.97
Part D Plan 2	478/677 (70.6%)	459/675 (68.0%)	0.93
Part D Plan 3	74/109 (67.9%)	72/109 (66.1%)	0.96
Part D Plan 4	49/71 (69.0%)	48/71 (67.6%)	0.97
Part D Plan 5	49/63 (77.8%)	48/63 (76.2%)	0.95

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

State Reliability

All state-level reliability scores were > 0.9; indicating that the measure would produce reliable scores at the state level.

Physician Group Reliability

The original denominator threshold tested was 30 patients, resulting in 53.5% (N=38) of 71 physician groups attributed having reliable scores (defined as 0.7 or greater). Increasing the denominator size to 45 patients resulted in 94.4% (N=34) of 36 physician groups with a reliable score. Among these groups, overall reliability was 0.71, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between physician groups. Therefore, these results suggest that physician groups with 45 patients or more will produce reliable scores.

Health Plan Reliability

Results obtained by the final two independent programmers met the Kappa threshold of 0.9, and no further refinement of measure specification was deemed necessary.

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.

Empirical validity testing is not available for this measure at the time of this maintenance review. Analysis was not possible in the timeframe from NQF publication of this new evaluation criteria (September 2017) and submission of the testing form (January 2018). On March 9, 2018, the measure steward, CMS, met with NQF to

discuss submission of this measure. NQF requires empiric validity testing at the time of maintenance; however, they recognize the limitations of the timeframe for submission. NQF, CMS, and the contract team agreed that in leu of providing results of testing, it would be suitable to include a detailed plan for testing empiric validity before the next maintenance submission.

We will test measure performance score validity by examining correlations with meaningful measures of a similar quality construct (convergent validity) using the Spearman's rank correlation coefficient. We will analyze the convergent validity of the measures, evaluating the extent to which the measures *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) and *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) correlate. We hypothesize that health plans and provider groups that perform well at helping individuals with schizophrenia remain adherent to antipsychotic medications will also perform well at helping individuals with bipolar I disorder remain adherent to mood stabilizers. Both measures are indicators of overall quality of care for individuals with serious mental illness and should be correlated.

For health plan level testing, we will evaluate the correlation between *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) and *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) using Medicare-Medicaid Plan (MMP) encounter data. We will begin our initial testing using data already available to us from federal fiscal years 2015 and 2016, covering dates between October 1, 2014, and September 30, 2016. Because of the uncertain quality of the encounter data reported by MMPs, we will conduct an initial series of data checks to examine the quality and volume of encounter data required for the measures and include MMPs for which the quality is sufficient for testing purposes. Our initial data checks will examine quality and volume of data at the plan and state levels to ensure sufficient sample sizes for testing the research questions. We anticipate using data elements related to Medicare and Medicaid enrollment, institutional encounters, non-institutional encounters, and prescription drug coverage and claims.

For provider level testing we will evaluate the correlation between *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) and *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) using Medicare FFS data paired with Medicare Part D claims data. We will pull this Medicare FFS data from the Integrated Data Repository (IDR) to complete testing. No Medicaid data will be used. We anticipate using data elements related to Medicare and Medicaid enrollment, institutional claims, non-institutional claims, and prescription drug coverage and claims.

We will produce scatter plots comparing the two measures at the provider and health plan level. The Spearman's rank correlation coefficient (r_s) assesses the monotonic relationship in plan rankings for each measure pair. The coefficient ranges from -1 to 1, where $r_s = 1$ indicates perfect alignment of plan rankings, $r_s = -1$ indicates opposite alignment of plan rankings, and $r_s = 0$ represents no alignment in plan rankings. We will fit a smooth curve using locally weighted scatterplot smoothing (LOWESS) method to visualize any trends in the scatterplots. Because the LOWESS method does not rely on a preconceived model for the distribution of the measures (non-parametric), the LOWESS curve can captured detailed information about the measure relationships that the correlation coefficient does not convey.

The timeline for this work is described below:

- October – November 2018: Develop analytic file
- November 2018 – February 2019: Conduct validity testing and review results
- March – April 2019: Summarize results and update measure documentation
- TBD: Submit updated validity testing to NQF as part of maintenance submission

Although empiric validity analysis has not yet been conducted, this measure uses a definition of adherence (0.8 proportion of days covered) that is harmonized with other National Quality Forum (NQF) endorsed adherence measures and is consistent with the threshold of adherence used in the seven studies cited in the evidence attachment. These studies demonstrated improved outcomes in schizophrenia associated with adherence to medication. Although many of these studies have used the medication possession ratio (MPR) rather than the

proportion of days covered (PDC), CMS and the Pharmacy Quality Alliance (PQA) have evaluated and extensively tested the PDC and the MPR and specifically found that: 1) the PDC and MPR will provide nearly identical results when examining adherence to a single drug; 2) the PDC will provide a more conservative estimate of adherence when examining adherence to a class of drugs that are prone to frequent switching and concomitant therapy with multiple drugs within the class (as with antipsychotic drugs). Therefore, based on NQF's recommendation that a standard methodology for calculating medication adherence be established across all endorsed adherence measures, CMS and PQA agreed to harmonize the methodology for calculating medication adherence using the PDC, which was approved by the NQF Consensus Standards Approval Committee (CSAC).

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

A Technical Expert Panel (TEP), comprising internal medicine physicians and pharmacists, evaluated the face validity of the measure and measure scores. The following 12 TEP members evaluated the face validity of the measure and measure scores:

1. Jill S. Borchert, Pharm.D., BCPS, Professor, Pharmacy Practice & PGY1 Residency Program Director, Midwestern University, Chicago College of Pharmacy
2. Anne Burns, RPh, Vice President, Professional Affairs, American Pharmacists Association
3. Jannet Carmichael, Pharm.D., FCCP, FAPhA, BCPS, VISN 21 Pharmacy Executive, VA Sierra Pacific Network
4. Marshall H. Chin, MD, MPH, Professor of Medicine, University of Chicago
5. Jay A. Gold, MD, JD, MPH, Senior Vice President and Medicare Chief Medical Officer, MetaStar, Inc.
6. David Nau, Ph.D., R.Ph., CPHQ, Senior Director of Research & Performance Measurement, PQA, Inc.
7. N. Lee Rucker, M.S.P.H., Senior Strategic Policy Advisor, AARP - Public Policy Institute
8. Marissa Schlaifer, MS, RPh, Director of Pharmacy Affairs Academy of Managed Care Pharmacy
9. Brad Tice, Pharm.D., Chief Clinical Officer, PharmMD Solutions, LLC
10. Jennifer K. Thomas, Pharm.D., Manager, Pharmacy Services, Delmarva Foundation for Medical Care/Delmarva Foundation of the District of Columbia
11. Darren Triller, Pharm.D., Director, Pharmacy Services, IPRO
12. Neil Wenger, MD, Professor of Medicine, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research

The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using Survey Monkey). Face validity of the measure score as an indicator of quality was systematically assessed as follows: After the measure was fully specified and tested, the expert panel members were asked to rate, based on a 5-point Likert scale, their level of agreement with the following statement: "The measure appears to measure what is intended."

The 5-point Likert scale was defined as follows: 1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree

ICD-10-CM Conversion Methodology

The conversion of the measure to include ICD-10-CM codes is provided as requested by NQF. The crosswalk is provided as an excel file in Section S2.b Data Dictionary or Code Table.

Name and Credentials of Experts Who Assisted in the Process

- Soeren Mattke, MD, DSc, Senior Scientist, RAND Corporation

- Tim Laios, MBA, MPH, Executive Director, Informatics, Health Services Advisory Group (HSAG)
- Ryan Fair, BS, Director, Informatics, HSAG
- Kerri Carlile, MS, Informatics Analyst, HSAG
- Sara Lomeli, BA, Informatics Project Coordinator, HSAG

Evaluation of ICD-9-CM Changes
The changes (i.e., deletions and/or additions) made to the ICD-9-CM codes for the measures requiring conversion were reviewed. Additionally, the ICD-9-CM codes were reviewed for any coding updates, using the October 2011 Conversion Table of New ICD-9-CM Codes, published by the National Center for Health Statistics (NCHS) and the Centers for Medicare & Medicaid Services (CMS).

ICD-9-CM Code Identification
For each measure requiring conversion, original tables were used to identify all ICD-9-CM codes included in the measure. Those ICD-9-CM codes and matching descriptions were then extracted from the Ingenix 2011 ICD-9-CM Data File. Only valid ICD-9-CM codes were retained and used in the ICD-9-CM to ICD-10-CM conversion process.

Ingenix Extraction
When extracting the ICD-9-CM codes from the Ingenix Data File, all codes were extracted with two-decimal specificity. For example, for ICD-9-CM code 274.1, all ICD-9-CM codes that had 2741 for the first four digits were extracted (e.g., 274.10, 274.11, and 274.19). For every three-digit ICD-9-CM code used in the measure, all ICD-9-CM codes that began with those first three digits were extracted. For the ICD-9-CM codes listed in ranges, codes with up to two-decimal specificity were extracted within that range.

Conversion Process
The ICD-9-CM and ICD-10-CM General Equivalence Map (GEM) text files and the ICD-10-CM Descriptions text file were imported into SAS and combined into one data file to capture all ICD-9-CM codes, their corresponding ICD-10-CM codes, and the ICD-10-CM code descriptions. The ICD-9-CM codes that were retained from the Ingenix 2011 ICD-9-CM Data File described above were then extracted from the combined GEM data file.

The results for each measure were then exported into Excel and validated to ensure that the translation was appropriate (i.e., the crosswalk was correct and applied appropriately and all appropriate ICD-9-CM codes were captured). Since one ICD-9-CM code can have several corresponding ICD-10-CM codes, each ICD-9-CM code can have multiple entries in the final Excel document (i.e., one row for each corresponding ICD-10-CM code).

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

Systematic Assessment of Face Validity
The results of the Technical Expert Panel rating of face validity as represented by this statement, "The measure appears to measure what is intended," on a scale of 1 to 5.
N=12 panel members, Mean Rating=4.33

Response / % of TEP / Number of TEP
Strongly Agree / 33.3% / 4
Agree / 66.7% / 8
Neutral / 0.0% / 0
Disagree / 0.0% / 0
Strongly Disagree / 0.0% / 0

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?*)

In summary, 100% of the TEP members responded "agree" or "strongly agree" with the statement that the measure, as specified, had face validity. The results indicate strong support of the face validity of the measure by the Technical Expert Panel.

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*)

Type of analysis

A sensitivity analysis was conducted to estimate the effect of the exclusion on the overall measure rate across the eight-state sample. The overall prevalence of the exclusion was calculated and the measure rate was calculated two ways: 1) with the exclusion applied and 2) without the exclusion applied.

Description of exclusion

Individuals with a diagnosis of dementia were excluded from the measure denominator. In April 2005, the Food and Drug Administration (FDA) issued a Public Health Advisory, which warned of the increased risk of mortality associated with the use of atypical antipsychotics in elderly patients with dementia. This warning was based on the findings of a meta-analysis of 17 short-term, randomized, placebo-controlled trials and showed that the risk of death in drug-treated patients was 1.6 to 1.7 times the risk of death in placebo-treated patients (Schneider et al., 2005). In 2008, the FDA Advisory and Black Box Warning was extended to all antipsychotic medications when further studies (Liperoti et al., 2009; Schneeweiss et al., 2007; Setoguchi et al., 2008) showed that conventional antipsychotics were associated with a similar increased risk of death when administered to elderly patients with a diagnosis of dementia.

Liperoti, R., Onder, G., Landi, F., Lapane, K. L., Mor, V., Bernabei, R., & Gambassi, G. (2009). All-cause mortality associated with atypical and conventional antipsychotics among nursing home residents with dementia: A retrospective cohort study. Journal of Clinical Psychiatry, 70(10),1340-1347.

Schneeweiss, S., Setoguchi, S., Brookhart, A., Dormuth, C., & Wang, P. S. (2007). Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ, 176, 627–632. [PubMed: 17325327]

Schneider, L. S., Dagerman, K. S., & Insel, P. (2005). Risk of death with atypical antipsychotic drug treatment for dementia: Meta-analysis of randomized placebo-controlled trials. Journal of the American Medical Association, 294, 1934–1943. [PubMed: 16234500]

Setoguchi, S., Wang, P. S., Brookhart, M., Canning, C. F., Kaci, L., & Schneeweiss, S. (2008). Potential causes of higher mortality in elderly users of conventional and atypical antipsychotic medications. JAGS, 56, 1644–1650.

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)

Individuals with dementia represented approximately 11% of all individuals in the measure denominator. If individuals with dementia were excluded, the measure rate was 74.4% (31,752/42,676) across the eight-state sample; whereas, the measure rate without excluding these individuals was 74.0% (35,416/47,852).

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*)

While overall performance across the eight-state sample did not differ, individuals with dementia represent a population where adherence to antipsychotic medications is associated with an increased risk of mortality. Therefore, the Technical Expert Panel recommended excluding this subpopulation.

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** 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)*

To identify statistically significant differences in performance for states and physician groups, we conducted a comparison of means and percentiles. Confidence intervals (95%) were calculated around point estimates and then compared to the grand mean of states. If the confidence intervals did not overlap with the overall grand mean, the comparison was considered statistically significant.

For physician groups and health plans, the observed sample sizes of members of each comparison unit were tested empirically to determine whether there was sufficient power to detect statistically significant differences between members (e.g., between plans or between physician groups). To do this, all members were divided into quintiles according to their measure score. Within each quintile, the member with a denominator closest in size to the median denominator of the quintile and the member with the measure score closest to the median measure score of that quintile were identified. Comparison of members based on their median denominator size was made, because a relationship between denominator size and quality cannot be excluded a priori. In addition, a "standardized" member of each quintile was simulated by using the median denominator size across all quintiles. Binomial (exact) 95% confidence intervals for each of the 10 selected plans or physician groups (i.e., two plans or physician groups per quintile) were calculated around the point estimates. Overlapping confidence intervals indicate insufficient statistical power to detect statistically significant differences.

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*)

Meaningful Differences at the State Level

Below we present the measure rate by state, mean, median, and standard deviation.

State A – 67.5%* (statistically significantly lower than the mean)
State B – 76.4%
State C – 71.0%* (statistically significantly lower than the mean)
State D - 84.7%* (statistically significantly higher than the mean)
State E – 80.0%
State F - 68.5%* (statistically significantly lower than the mean)
State G - 78.3%
State H - 81.1%
Mean of state scores – 75.9%
Median of state scores – 77.4%
Standard deviation of state scores – 6.3%

Meaningful Differences at the Physician Group Level

Below we present the mean, standard deviation, and percentiles at the physician group level.
Number of Physician Groups with at least 45 individuals in measure denominator = 36
Mean: 81.4%
SD: 10.8%
10th Percentile: 68.0%
25th Percentile: 77.9%
50th Percentile: 82.6%
75th Percentile: 89.0%
90th Percentile: 92.3%
Of physician group scores, 8.3% were statistically significantly lower than the mean, and 33.3% of physician group scores were statistically significantly higher than the mean, indicating a wide range of scores.

Across Physician Groups with ≥ 30 Beneficiaries	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Number of physician groups	6	5	6	6	5
Denominator range across physician groups (minimum-maximum)	30-140	30-37	31-73	39-143	30-46
Median denominator size	50	34	37	55	42
Measure score (95% CI) of the physician group with a denominator size closest to the	66.7% (53.9-80.0)	73.5% (60.0-87.1)	75.8% (62.5-88.9)	81.5% (71.9-90.7)	88.1% (79.6-96.0)

Across Physician Groups with ≥ 30 Beneficiaries	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
median denominator size					
Measure score range across physician groups (minimum-maximum)	37.9%-66.7%	69.7%-73.5%	75.0%-77.6%	79.5%-83.6%	85.7%-93.5%
Median measure score	61.2%	73.0%	77.1%	81.7%	88.1%
Measure score (95% CI) of the group with a score closest to the median score	63.1% (54.2-72.4)	73.0% (59.9-86.2)	77.4% (64.2-90.4)	81.5% (71.9-90.7)	88.1% (79.6-96.0)
95% CI using the overall median denominator N=42	61.2% (47.5-75.8)	73.0% (60.6-85.5)	77.1% (65.4-88.6)	81.7% (71.0-91.9)	88.1% (79.6-96.0)
CI = Confidence Interval					

Meaningful Differences at the Health Plan Level

Across Part D Plan with ≥ 30 Beneficiaries	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Number of plans	4	5	5	5	4
Denominator range across plans (minimum-maximum)	34-220	97-1,267	238-3,188	792-3,304	53-413
Median denominator size	117	314	2,234	1,338	212
Measure score (95% CI) of the plan with a denominator size closest to the median denominator size	63.3% (55.6-71.3)	67.8% (62.8-73.0)	70.5% (68.7-72.4)	74.0% (71.7-76.3)	78.5% (73.1-83.8)
Measure score range across plans	58.8%-64.1%	66.0%-67.8%	69.4%-71.4%	72.5%-75.6%	77.4%-81.8%
Median measure score	63.6%	66.1%	69.7%	74.0%	78.0%
Measure score and 95% CI of the plan with a score closest	63.3% (55.6-71.3)	66.1% (61.5-70.8)	69.7% (64.1-75.5)	74.0% (71.7-76.3)	78.5% (73.1-83.8)

Across Part D Plan with ≥ 30 Beneficiaries	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
to the median score					
95% CI based on the overall median denominator size N=389	63.6% (58.9-68.4)	66.1% (61.5-70.8)	69.7% (65.3-74.3)	74.0% (69.7-78.3)	78.0% (74.0-82.0)
CI = Confidence Interval					

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?*)

Meaningful Differences at the State Level
Three states (37.5%) had scores statistically significantly lower than the mean and one state (12.5%) had scores significantly higher than the mean. Measure rates by state ranged from 67.5% in state A to 84.7% in state D, indicating suboptimal performance across all states and variation between high- and low-performing states.

Meaningful Differences at the Physician Group Level
The testing results indicate ample room for improvement and meaningful differences in quality of care between the highest and lowest performing physician groups. Overall 41.6% of physician performance scores were statistically different from the mean. For those physician groups with at least 45 eligible individuals, high- (90th percentile) and low- (10th percentile) performing physician groups were 24.3 percentage points apart.

Please note after testing was conducted the measure was harmonized to include individuals receiving depot injections (rather than exclude those individuals). The testing data presented above do not yet reflect the change in specification. Our preliminary testing since the addition of individuals receiving depot injections showed that the impact of this inclusion increases the denominator size by approximately 23% and decreases the overall measure rate across the eight-state sample by 2.2 percentage points.

A total of 28 physician groups with at least 30 beneficiaries were identified and could be distributed across the measure score quintiles. Physician groups showed limited variation in sample size with no particular pattern with respect to measure scores. We noted pronounced variation in measure rates across physician groups, ranging from 37.9% to 93.5%, but denominator sizes were consistently small, resulting in wide confidence intervals. Comparison of standardized physician groups (calculated based on the score closest to the median measure score or the overall median denominator size) showed sufficient discriminatory ability between physician groups of the highest and lowest quintiles.

Assuming a median measure rate of 77.1% and a median denominator of 42 beneficiaries, the smallest statistically significant difference that can be detected at the physician group level with a power of 80% and α=0.05 is 18.0%.

Meaningful Differences at the Health Plan Level
A total of 23 plans with at least 30 beneficiaries could be distributed across the measure score quintiles. Plans showed pronounced variation in sample size with a general pattern in the first 4 quintiles of increasing size with respect to measure scores. Comparison of individual plans (selected based on the denominator size closest to the median denominator size or score closest to the median measure score) showed sufficient discriminatory ability, based on lack of overlap between the confidence intervals of the lowest and highest performing quintiles and limited discriminatory

ability between the lowest quintile and the 4th quintile. Comparison of standardized plans (with confidence intervals calculated based on the overall median denominator size of the entire sample) showed sufficient discriminatory ability between members of the highest and lowest quintiles, as well as between the lowest quintile and the 4th quintile. Of note, the sample sizes for plans varied dramatically within each quintile and will result in distinctly different power if two members are compared.

Assuming a median measure rate of 69.7% and a median denominator of 389 beneficiaries, the smallest statistically significant difference in measure rates that can be detected at the plan level with a power of 80% and α=0.05 is 6.9%.

Please note after testing was conducted the measure was harmonized to include individuals receiving depot injections (rather than exclude those individuals). The testing data presented above do not yet reflect the change in specification. Our preliminary testing since the addition of individuals receiving depot injections showed that the impact of this inclusion increases the denominator size by approximately 23% and decreases the overall measure rate across the eight-state sample by 2.2 percentage points.

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 days' supply data and bias from cash prescriptions were possible threats to validity. An empirical assessment of these possible threats was conducted as follows:

Threat of Bias from Missing Data
We have identified two potential scenarios where measure results could be biased by missing data:

1. Missing days' supply within the prescription drug event data, which is a required data element to calculate medication adherence;
2. Missing drug claims due to individuals using alternative payment sources for prescription drugs, e.g., $4 commercial discount prescription programs and other alternative drug benefits, such as the Veterans Administration (VA)

For missing days' supply, we analyzed the number (%) of beneficiaries in the measure denominator with one or more claims that had missing days' supply.

For bias from cash prescriptions or alternative sources, we conducted a limited sensitivity analysis using a two-state sample (states C and G) to estimate the potential impact of a commercial cash discount program on measure rates. Specifically, we created a National Drug Code (NDC) list from the formulary of a leading cash discount program to identify those individuals with at least one claim for an antipsychotic on the formulary and no claims for any other Part D drugs on the formulary as a proxy to potentially indicate the individual was filling medications through the cash discount program. We then simulated the effect on measure rates, if each of these individuals' antipsychotic drug use extended from the last consecutive claim to the end of the measurement period, assuming that individuals had switched to the cash program. We simulated two scenarios: including complete coverage of all remaining days' until the end of the measurement period were 100% or extrapolating the average proportion of days covered from the first prescription in the measurement period to the last prescription in the measurement period.

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*)

Missing Data
Days' Supply: Only 2 individuals (0.005%) in the overall measure denominator had one or more claims with missing days' supply.

Cash Prescriptions
The percentage of individuals in the denominator with antipsychotic Part D claims on the formulary and no claims for any other drugs on the commercial discount formulary was 0.9% (145/15,874).

SCENARIO 1. If individuals with possible cash prescriptions (i.e., those described above) are assumed to have antipsychotic medication for all days from the last day covered to the end of the measurement period (i.e., 100% adherence), the PDC would be 71.6% (11,365/15,874).

SCENARIO 2. If individuals with possible cash prescriptions (i.e., those described above) are assumed to have antipsychotic medication for all days from the last day covered at the same proportion as the PDC calculated over the period from first to last claim in the measurement period (i.e., same adherence as the rest of the period), the PDC would be 71.5% (11,353/15,874).

The actual measure rate was 71.5% (11,348/15,874).

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

Only 2 individuals (0.005%) in the overall measure denominator had one or more claims with missing days' supply. This small number indicates that missing data do not pose a threat to the validity of the measure.

Cash Prescriptions

The actual measure rate was 71.5% (11,348/15,874). Therefore, the findings suggest that very little impact on measure rates would be expected from utilization of the cash discount program. In addition, since the most prevalent antipsychotic medications are not included in the commercial discount program due to their cost, it is unlikely that commercial discount programs will have an impact on measure rates in the near-term. Of note, this analysis is exploratory in nature and assumes that individuals were not switched to a drug on the commercial discount formulary, and if they were utilizing the discount program, they were obtaining all of their medications at a cash discount program. Additional limitations include prescriptions filled with other benefits (e.g., VA), and the extent to which this measure might underestimate antipsychotic use due to those factors is unknown.

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.
Testing demonstrated that the data required were available and accessible. Issues affecting feasibility regarding missing data were not identified. The cost of data collection is negligible, since the administrative data (collected by CMS primarily for billing purposes) are used as the data source for this measure. Other feasibility/implementation issues were not identified.

Eligible professionals successfully reported this measure to CMS as part of the Physician Quality Reporting Program.

DATA COLLECTION
Testing was conducted with the CMS administrative claims data. No additional data collection was conducted.

AVAILABLILITY OF DATA
Testing was conducted with the CMS administrative claims data. The data were readily available and accessible.

MISSING DATA
No threats to the validity of this measure were identified using a limited analysis designed to address missing data (Reference Validity Testing Section 2b2.2).

TIMING AND FREQUENCY OF DATA COLLECTION
Testing was conducted with the CMS administrative claims data. Data sources needed to implement the measure are collected by CMS in a timely manner.

SAMPLING
Not Applicable

PATIENT CONFIDENTIALITY
Not Applicable

TIME AND COST OF DATA COLLECTION
The administrative data (collected by CMS primarily for billing purposes) are used as the data source for this measure. Therefore, the cost of data collection is negligible.

OTHER FEASIBLITY/IMPLEMENTATION ISSUES
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)*.
Proprietary coding is contained in the attached list of codes. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets.

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 (Internal to the specific organization)	Public Reporting

Not in use	Physician Compare https://www.medicare.gov/physiciancompare/ Payment Program Quality Payment Program (previously PQRS) https://qpp.cms.gov/mips/quality-measures New York State Delivery System Reform Incentive Payment (DSRIP) Program https://www.health.ny.gov/health_care/medicaid/redesign/dsrip/vbp_library/quality_measures/docs/2018_harp_qms.pdf Quality Improvement (external benchmarking to organizations) Substance Abuse and Mental Health Services Administration (SAMHSA) section 223 demonstration https://www.samhsa.gov/sites/default/files/programs_campaigns/ccbhc-criteria.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

Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): This measure is used in the Quality Payment Program (QPP) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible clinicians. Quality performance results from QPP will be published on Physician Compare.

New York State Delivery System Reform Incentive Payment (DSRIP) Program: The measure is publicly reported in New York State's Delivery System Reform Incentive Payment (DSRIP) Program, and is included in the Value Based Payment (VBP) Quality Measure Set for the Health and Recovery Plan (HARP) subpopulation. As of 2016, 45,000 individuals were enrolled in HARP. HARP is a specialized managed care program for adult individuals with Severe Mental Illness (SMI) or Substance Use Disorder (SUD) that began its rollout in New York State on October 1, 2015. This measure was selected as clinically relevant, reliable, valid, and feasible and is required to report. Pay for performance measures are intended to be used in the determination of shared savings amount for which VBP Contractors are eligible. In other words, these are the measures on which payments in VBP contracts may be based. Measures can be included in both the determination of the target budget and in the calculation of shared savings for VBP Contractors.

Substance Abuse and Mental Health Services Administration (SAMHSA) Section 223 Demonstration Program: This program is authorized under Section 223 of the Protecting Access to Medicare Act (PAMA). Program activities aim to integrate behavioral health with physical health care, increase consistent use of evidence-based practices, and improve access to high-quality care. Participating states in the demonstration program certify community behavioral health clinics that meet federally developed criteria emphasizing accessible and high-quality care. The certified community behavioral health clinics (CCBHCs) are compensated for services through a prospective payment system (PPS).

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?*)

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.*)

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.
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): In 2015, 80 eligible professionals (EP) reported on the measure. EPs submitting PQRS data to CMS received a PQRS feedback report on whether they satisfactorily reported and if they are subject to a payment adjustment. The data in these reports may help EPs determine whether or not it is necessary to submit an informal review request. An informal review is a process that allows EPs to request a review of their payment adjustment determination.

New York State DSRIP Program: This measure was added to the program to be tested in the HARP subpopulation in 2017 with results to be reported in 2018. Medicaid Managed Care Organizations with Level 1 or higher value–based contracting arrangements or MCOs with a VBP Pilot contract are required to report. The New York State Department of Health website provides a library of resources for providers and health plans including the technical specifications manual, webinars, and information about the advisory groups involved. The state also holds workshops to explain the VBP process and expectations.

SAMHSA Section 223 Demonstration Program: In 2015, the Department of Health and Human Services (HHS) awarded CCBHC planning grants (Phase I) to 24 states, and eight of those states were selected to participate in the demonstration program (Phase II) to improve access to high-quality behavioral health programs. The CCBHC demonstration program and PPS are designed to work within the scope of state Medicaid Plans and to apply specifically to individuals who are Medicaid enrollees. The eligible population in these states includes all behavioral health clinic (BHC) consumers served by a BHC provider.

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.**
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): Each year, QPP individual EPs and QPP group practices receive feedback reports on whether they satisfactorily reported and if they are subject to the future downward payment adjustment. CMS hosts training sessions on these reports and posts audio recording and slide presentations on their webpages. CMS also provides technical assistance and maintains webpages with information about accessing and understanding these reports.

New York State DSRIP Program: Information on the process are provided in New York State's, 2018 Value Based Payment Reporting Requirements Technical Specifications Manual. Plans will electronically submit patient-level detail files and patient attribution files via secure file transfer on August 1, 2018. The New York State Department of Health website provides a library of resources for providers and health plans including the technical specifications manual, webinars, and information about the advisory groups involved. The state also holds workshops to explain the VBP process and expectations.

SAMHSA Section 223 Demonstration Program: Certified community behavioral health clinics and their states are required to collect 21 of 32 quality measures for the demonstration program. This measure is required to be reported. For each demonstration year (the measurement year), quality measures and metrics are submitted within nine months for CCBHCs, and within 12 months for states. CCBHC-lead data and measures are reported to their designated state agency, and state-lead data and measures are reported to SMAHSA by email. SAMHSA will share the data with CMS for the purposes of Quality Bonus Payments and with the Office of the Assistant Secretary for Planning and Evaluation (ASPE) for the purposes of evaluation. Data is reported by using the data reporting templates, and relaying on the major specifications and instructions for those templates found in the Technical Specifications and Resource Manual. SAMHSA's technical assistance (e.g. webinars, guidance documents) is designed to help states and clinics collect, analyze and report the data for each measure. Clarifications related to quality measures and data reporting are provided on the SAMHSA website, and additional questions are submitted by email.

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.
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): CMS solicits feedback and has a designated space on their webpage with information on how to share feedback with them. The measure owner has not received any feedback on this measure.

New York State DSRIP Program: The program is in its first pilot year and performance has not yet been reported. The state receives feedback on quality measure feasibility, reporting, and calculation from a VBP Measure Support Task Force, including professionals from various Managed Care Organizations (MCOs), VBP Pilot Contractors, State Agencies, along with other professionals with experience in quality measurement and health information technology. They also receive input from a Clinical Advisory Group that evaluates feedback from VBP Contractors, MCOs, and stakeholders, any significant changes in evidence base

of underlying measures and/or conceptual gaps in the measurement program. Feedback from these groups is not publicly available at this time.

SAMHSA Section 223 Demonstration Program: For the purposes of continuous quality improvement, behavioral health clinics (BHCs) submit data and measure results to the state. Ongoing refinement of the system at both the state and BHC level is achieved through state feedback to the BHC regarding the data and measure results, and BHC internal feedback and adjustment regarding both data and results. Feedback from these groups is not publicly available at this time.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): No feedback was received specific to this measure.

New York State DSRIP Program: No feedback specific to this measure is currently available.

SAMHSA Section 223 Demonstration Program: No feedback specific to this measure is currently available.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure recently went through a re-evaluation process. During that process, feedback on the measure was obtained from measure advisory panels including NCQA's Pharmacy Panel and NCQA's Behavioral Health Measure Advisory Panel. These panels recommended adding medications which are FDA approved for the treatment of schizophrenia and removing medications which are not FDA approved.

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.**
Based on the feedback obtained from NCQA's Pharmacy Panel and Behavioral Health Measure Advisory Panel (described in 4a2.2.3) the following measure changes were implemented:
1. Add the following FDA approved medications to the measure:
- Cariprazine
- Quetiapine fumarate (Seroquel)
- Brexpiprazole
- Aripiprazole lauroxil (Aristada)

2. Remove the following off-label medications from the measure (these medications were included in the original measure specification):
- Pimozide
- Olanzapine-fluoxetine

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.
Quality Payment Program (QPP) - previously Physician Quality Reporting System (PQRS): PQRS data extracted from Physician Compare is only available for 2015. Data was not available at the time of maintenance endorsement to evaluate improvement. In future endorsement maintenance we will be able to show change over time and hope to demonstrate improvement in performance.

New York State DSRIP Program: Performance data is not publicly available for this measure.

SAMHSA Section 223 Demonstration Program: Performance data is not publicly available for this measure.

We envision this measure will help providers to identify patients with schizophrenia who are not adherent (at a critical threshold of 0.8 or greater) with long-term treatment with antipsychotic medications and target interventions to improve medication adherence.

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.**

Susceptibility to inaccuracies, errors, or unintended consequences were not identified during testing. There were no identified unintended findings for this measure during testing and none have been brought to our attention since implementation.

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

No unexpected benefits.

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)**

0541 : Proportion of Days Covered (PDC): 3 Rates by Therapeutic Category
0542 : Adherence to Chronic Medications
0543 : Adherence to Statin Therapy for Individuals with Cardiovascular Disease
0544 : Use and Adherence to Antipsychotics among members with Schizophrenia
0545 : Adherence to Statins for Individuals with Diabetes Mellitus
0569 : ADHERENCE TO STATINS
1880 : Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder

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

Adherence to Antipsychotic Medications for Individuals with Schizophrenia. NCQA is measure 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 specifications are harmonized with the related measure, Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder (NQF #1880), where possible. The methodology used to calculate adherence in these measures is proportion of days covered (PDC) which is calculated the same in both measures. The methodology used to identify the denominator population is

also calculated the same in both measures with the exception of the clinical conditions which is the target of the measure. The medications included in both measures are specific to the clinical condition targeted in the measure.

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 Adherence to Antipsychotic Medications for Individuals with Schizophrenia (NCQA) measure is used for HEDIS reporting and is harmonized with the NQF #1879 in condition, target population, methodology, and medications. The HEDIS measure is only used in Medicaid health plans and therefore is restricted to adults age 18-64.

During development the measure developers identified another competing measure which eventually lost NQF endorsement. The section below is from the original submission of the measures for initial endorsement and compares this measure (#1879 Adherence to Antipsychotic Medications for Individuals with Schizophrenia) to a previously NQF-endorsed measure (#0544 Use and Adherence to Antipsychotics among Members with Schizophrenia).

Measure 1879 (Adherence to Antipsychotic Medications for Individuals with Schizophrenia) has both the same measure focus and essentially the same target population as Measure 0544 (Use and Adherence to Antipsychotics among Members with Schizophrenia), which is no longer endorsed after the measure's time-limited endorsement (TLE) status expired. Measure 1879 is superior to the existing Measure 0544 because it represents a more valid and efficient approach to measuring medication adherence to antipsychotic medications. In addition, as discussed above in Section 5a.2, Measure 1879 is harmonized with several other adherence measures in the NQF portfolio. Key differences in measure validity and efficiency are addressed in the sections below.

VALIDITY
The Proportion of Days Covered (PDC), which is the method used to calculate adherence in Measure 1879, has several advantages over the Medication Possession Ratio (MPR), which is used in Measure 0544. First, the PDC was found to be more conservative compared to the Medication Possession Ratio (MPR) and was preferred in clinical scenarios in which there is the potential for more than one drug to be used within a drug class concomitantly (e.g., antipsychotics). This clinical situation applies directly to Measure 1879. Martin et al. (2009) demonstrated this in a study published in the Annals of Pharmacotherapy by comparing the methodology for drugs that are commonly switched, where the MPR was 0.690, truncated MPR was 0.624, and PDC was 0.562 and found significant differences between the values for adherence ($p < 0.001$). Martin et al (2009) also compared drugs with therapeutic duplication where the PDC was 0.669, truncated MPR was 0.774, and MPR was 1.238, and again obtained significant differences ($p < 0.001$). These findings were partially replicated by testing results from FMQAI (now HSAG) of Measure 1879 where MPR produced a higher measure rate (as compared to PDC) as shown below.

Adherence to Antipsychotic Medications for Individuals with Schizophrenia
Method Measure Rate

Comparison of MPR and PDC
Method Measure Rate
MPR 74.4%
PDC 70.0%
Based on initial draft measure specifications and data from a 100% sample of Medicare fee-for-service beneficiaries with Part D coverage in Florida and Rhode Island, using 2008 Medicare Parts A, B, and D data.

Additional differences between Measure 1879 and TLE 0544 related to validity include the following concerns:

Denominator: The measure denominator requires at least two antipsychotic medication prescriptions; whereas, the NQF TLE measure (NQF# 0544) does not require any antipsychotic medication prescriptions in the measure denominator. In 0544, an MPR

of "0" is assigned to those without any antipsychotic medication prescriptions, which may falsely lower measure rates, specifically in scenarios where the prescriber has made the decision not to prescribe antipsychotic medications for an individual diagnosed with schizophrenia.

Exclusion related to a diagnosis of dementia: Measure 1879 excludes individuals with a diagnosis of dementia during the measurement year which is not considered in Measure 0544. Antipsychotic medications are currently labeled with a Food and Drug Administration (FDA) Black Box warning that states, "Elderly patients with dementia-related psychosis treated with antipsychotic drugs are at an increased risk of death. Analyses of seventeen placebo-controlled trials (modal duration of 10 weeks), largely in patients taking atypical antipsychotic drugs, revealed a risk of death in drug-treated patients of between 1.6 to 1.7 times the risk of death in placebo-treated patients." The Technical Expert Panel, which reviewed the measure, recommended excluding these individuals from the measure denominator, since continued adherence to antipsychotic medications in this subpopulation may increase mortality and not represent quality of care. (Please see Section 2b3.2 that provides descriptive results of testing related to exclusions.)

EFFICIENCY
Measure 1879 requires only one year of administrative claims data, rather than two years of data which is required for TLE 0544. The Technical Expert Panel that reviewed Measure 1879 indicated that the burden of requiring two years of administrative claims data would not meaningfully modify measure rates and would potentially result in the unnecessary exclusion of individuals for which adherence should be assessed but for which only 1 year of claims data were available. Additional rationale for this TEP recommendation was related to an increased length of the continuous enrollment criteria to specify the measure use with two years of data. FMQAI's (now HSAG) empirical analysis of a related adherence measure (NQF 0542 – Adherence to Chronic Medications) using 2007 and 2008 Medicare Part D data for beneficiaries in Florida and Rhode Island validated this concern and indicated that approximately 10% of the eligible population would be excluded from the measure if the enrollment criteria required two years of administrative claims data as opposed to one year.

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
Co.2 **Point of Contact:** Elizabeth, Ricksecker, Elizabeth.Ricksecker@cms.hhs.gov, 410-786-6723-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, swift@ncqa.org, 202-955-5174-

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.
Behavioral Health Measure Advisory Panel (BHMAP) – advised on the measure re-evaluation:
1. Katherine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute
2. Christopher Dennis, MD, MBA, FAPA, Chief Behavioral Health Officer, Landmark Health
3. Ben Druss, MD, MPH, Professor, Emory University
4. Frank A. Ghinassi, PhD, ABPP, President and CEO, Rutgers University Behavioral Health Care
5. Connie Horgan, ScD, Professor and Director, Institute for Behavioral Health, Brandeis University
6. Laura Jacobus-Kantor, PhD, Chief, Quality, Evaluation and Performance, SAMHSA HHS
7. Jeffrey Meyerhoff, MD, National Medical Director for Medicare and Retirement, Optum Behavioral Solutions
8. Harold Pincus, MD, Professor and Vice Chair--Department of Psychiatry, College of Physicians and Surgeons, Co-Director, Irving Institute for Clinical and Translational Research, Columbia University, Director of Quality and Outcomes Research, New York –Presbyterian Hospital

9. Michael Schoenbaum, PhD, Senior Advisor for Mental Health Services, Epidemiology and Economics, National Institute of Mental Health
10. John Straus, MD, Medical Director Special Projects, Massachusetts Behavioral Health Partnership A Beacon Health Options Company
11. William Wood, MD, PhD, Manager, Medical Director Behavioral Health, Anthem, Inc.

HEDIS Expert Pharmacy Panel – advised on the measure re-evaluation:
1. Linda DeLaet, PharmD, Kaiser Permanente
2. Gerry Hobson, RPh, Cerner Multum
3. Chronis H. Manolis, RPh, UPMC Health Plan
4. Cathrine Misquitta, PharmD, MBA, BCPS, CGP, FCSHP, Health Net Pharmaceutical Services
5. Kevin Mark, MD, Wisconsin First, Inc.

FMQAI (now HSAG) TEP - advised on the original measure development and testing:
1. Douglas Bell, MD, Associate Professor in Residence, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research
2. Jill S. Borchert, Pharm.D., BCPS, Professor, Pharmacy Practice and PGY1 Residency Program Director, Midwestern University, Chicago College of Pharmacy
3. Anne Burns, RPh, Vice President, Professional Affairs, American Pharmacists Association
4. Jannet Carmichael, Pharm.D., FCCP, FAPhA, BCPS, VISN 21 Pharmacy Executive, VA Sierra Pacific Network
5. Marshall H. Chin, MD, MPH, Professor of Medicine, University of Chicago
6. Edward Eisenberg, MD, Vice President and Chief Medical Officer, Medicare, Medco Health Solutions
7. Jay A. Gold, MD, JD, MPH, Senior Vice President and Medicare Chief Medical Officer, MetaStar, Inc.
8. David Nau, Ph.D., R.Ph., CPHQ, Senior Director of Research and Performance Measurement, PQA, Inc.
9. N. Lee Rucker, M.S.P.H., Strategic Policy Senior Advisor, AARP - Public Policy Institute
10. Marissa Schlaifer, MS, RPh, Director of Pharmacy Affairs Academy of Managed Care Pharmacy
11. Brad Tice, Pharm.D., Chief Clinical Officer, PharmMD Solutions, LLC
12. Jennifer K. Thomas, Pharm.D., Manager, Pharmacy Services, Delmarva Foundation for Medical Care / Delmarva Foundation of the District of Columbia
13. Darren Triller, Pharm.D., Director, Pharmacy Services, IPRO
14. Neil Wenger, MD, Professor of Medicine, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 04, 2018
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 04, 2019

Ad.6 **Copyright statement:** Not Applicable, the measure is in the public domain.
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 1880
Measure Title: Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: Percentage of individuals at least 18 years of age as of the beginning of the measurement period with bipolar I disorder who had at least two prescription drug claims for mood stabilizer medications and had a Proportion of Days Covered (PDC) of at least 0.8 for mood stabilizer medications during the measurement period (12 consecutive months).
Developer Rationale: We envision several important benefits related to quality improvement with the implementation of this measure. Specifically, the measure will help providers to identify patients with bipolar I disorder who are not adherent (at a critical threshold of 0.8 or greater) with long-term treatment with mood stabilizer medications. Guidelines from the American Psychiatric Association (APA) and the National Institute for Clinical Excellence (NICE) emphasize the importance of treatment adherence and uninterrupted mood stabilizer medication regimens to prevent symptoms and relapse. Furthermore, this measure will encourage providers to develop interventions to improve adherence for this high-risk population. Improved medication adherence among individuals with bipolar I disorder would be expected to result in better control of the initial episode, the prevention of relapse to the initial episode, and the recurrence of new manic or depressive episodes, and as a result, lower mental health-related hospitalization rates and lower suicide rates. APA recommends that pharmacotherapy must be applied in ways that yield good tolerability and do not predispose the patient to nonadherence. Adoption of this performance measure has the potential to improve the quality of care for individuals with bipolar I disorder and, therefore, advance the quality of care in the area of mental health, a priority area identified by the National Priorities Partnership.

Numerator Statement: Individuals with bipolar I disorder who had at least two prescription drug claims for mood stabilizer medications and have a PDC of at least 0.8 for mood stabilizer medications.
Denominator Statement: Individuals at least 18 years of age as of the beginning of the measurement period with bipolar I disorder and at least two prescription drug claims for mood stabilizer medications during the measurement period (12 consecutive months).
Denominator Exclusions: Not Applicable

Measure Type: Process
Data Source: Claims
Level of Analysis: Clinician : Group/Practice, Health Plan, Integrated Delivery System, Population : Regional and State

Original Endorsement Date: Mar 04, 2014 **Most Recent Endorsement Date:** Mar 04, 2014

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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**

Summary of prior review in 2014

- The Steering Committee considered the measure important because it focuses on monitoring the initial treatment and medication adherence of patients with Bipolar I Disorder, which has a lifetime prevalence rate of 1-3.3 percent in the adult population in the US.
- Studies have recorded a wide variability of adherence rates for patients, and there are also age related discrepancies noted for medication adherence with adults 18 to 64 as opposed to 64 years and older.
- The evidence demonstrated that low adherence rates are associated with higher rates of recurrence and relapse, psychiatric hospitalizations and suicides.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provides a logic model outlining the process of identifying patients with Bipolar I Disorder who are not adherent to mood stabilizer medication treatment and the relationship to improved symptom control for those patients identified and a reduction in hospitalization.
- The developer included two clinical practice guideline recommendations:
  - National Institute for Clinical Excellence (2014), Bipolar Disorder: Assessment and Management. Thirty-six randomized control trials were included. **Guidelines do not provide independent grades to each recommendation.**
  - American Psychiatric Association (2004), Practice Guidelines for the Treatment for Patients with Bipolar Disorder. **Grades to the recommendation are I (Recommended with substantial clinical confidence) or II (Recommended with moderate clinical confidence).**

Questions for the Committee:

- *The evidence provided by the developer is updated and directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

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 identified performance gaps and wide variation in adherence to mood stabilizer medications with a PDC of 0.8 or greater among persons with bipolar I disorder across states, Part D Plans, Accountable Care Organizations (ACOs), and physician groups.
- Additional literature was cited in support of performance gap specific to eight studies demonstrating low rates (ranging from 16% to 76%) of adherence among individuals with bipolar I disorder who are prescribed mood stabilizer medications.

Disparities

- The developer analyzed 2007 and 2008 claims data for Medicare beneficiaries to demonstrate existing disparities in race and age. Adherence rates for mood stabilizing medication were lower among African American and Hispanic persons with bipolar disorder compared with White persons.

Questions for the Committee:

- o *Does the Committee have any specific questions on the information provided for gap in care or disparities?*
- o *Does the gap in care continue to warrant 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:

**No concerns.
**The evidence is sufficient for this measure. The measure is also part of clinical practice guidelines recommendations.
**This process measure with updated evidence since the last review bases their measure on solid well researched and endorsed guidelines that recommend reliably and consistently taking mood stabilizer medications helps prevent and treats Bipolar Affective Disorder Type I depressions and manic episodes. I agree that it's moderately strong.
**Process measure.

1b. Performance Gap

Comments:

**Gap continues to exist.
**There is a rather significant gap in this performance measure as reported by the developer. Non adherence rates for persons with Bipolar 1 disorder are relatively high.
**Evidence (both literature based and developer data) is moderately strong that a performance gap exists.
**+ performance gap--opportunity for improvement exists.

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:
- *The 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:
- *Does the Committee have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
- *Is the Committee satisfied with the developers empirical validity testing plan and timeline?*

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

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

Committee pre-evaluation comments

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

2a1. Reliability – Specifications
Comments:
**Data elements were clearly defined. No concerns.
**It's adequately reliable.

2a2. Reliability – Testing
Comments:
**No concerns.
**No concerns. Hi reliability reported: .9 Kappa. Measured at the health plan level.
**I'm satisfied with the reliability.
**No--moderate reliability.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:
**No concerns.
**Missing data does not pose a threat to validity with this measure.
**I'm satisfied as to validity.
**Moderate validity.

2b2-3. Other Threats to Validity

2b2. Exclusions
2b3. Risk Adjustment
Comments:
**No risk adjustment.
**There is evidence that results show disparities by race and age. I'm not sure I understand why this process measure is NOT risk adjusted.

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 defined fields in electronic claims and readily available and accessible
- CPT proprietary coding is contained in the measure.

Questions for the Committee:

- *Does the Committee have any concerns in regards to the feasibility of the measure?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility
Comments:
**Electronic claims
**This measure could easily be generated because data elements are in electronic claims.
**Feasible--testing conducted with CMS claims data.

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**

Accountability program details

- The measure is currently publicly reported (though not required) in the New York State Delivery System Reform Incentive Payment (DSRIP) Program, and is included in the Value Based Payment (VBP) Quality Measure Set for the Health and Recovery Plan (HARP) subpopulation.

- The measure is used (optional) in the Substance Abuse and Mental Health Services Administration (SAMHSA) Section 223 Demonstration Program.

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

- The developer received feedback on the measure from NCQA's Pharmacy Panel and NCQA's Behavioral Health Measure Advisory Panel. The Panels recommended adding FDA approved medications and removing medications that are not FDA approved. Based on the feedback received measure changes were implemented.
- The developer has not received feedback from the New York State DSRIP program. Performance has not been reported for the program, as it is in its first pilot year.
- Feedback is not publically available from the SAMHSA demonstration program.

Additional Feedback: N/A

***Questions for the Committee**:*

- *Has the measure been vetted in real-world settings by those being measured or others to the Committee's satisfaction?*

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

- None reported. Performance data is not publically available.

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 have been identified.

Potential harms

- None identified.

***Questions for the Committee**:*

o *In the absence of performance results can the Committee determine if the measure can 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:

**Currently publicly reported and used in accountability programs.
**The New York State Delivery System Reform Payment Program and SAMHSA currently use this performance measure.
**I agree that evidence for this is moderate to high.
**Data not publicly available.

4b1. Usability – Improvement

Comments:

**Concern: some meds without FDA approval may still prove effective for individuals with bipolar (typical antipsychotics, paliperidone...).
**This measure could be helpful in identifying individuals that are not adherent with their medication, thereby reducing potential decompensation and costly crisis care and hospitalizations.
**It's been used in the real world enough to say it's adequately usable.
**Tracking of this measure can help providers identify people with bipolar 1 disorder who are not adherent with their treatment.

Criterion 5: Related and Competing Measures

Related or competing measures

There are no competing measures. The developer includes the following related measures:

0003 : Bipolar Disorder: Assessment for diabetes
0109 : Bipolar Disorder and Major Depression: Assessment for Manic or hypomanic behaviors
0110 : Bipolar Disorder and Major Depression: Appraisal for alcohol or chemical substance use
0111 : Bipolar Disorder: Appraisal for risk of suicide
0112 : Bipolar Disorder: Level-of-function evaluation
0541 : Proportion of Days Covered (PDC): 3 Rates by Therapeutic Category
0542 : Adherence to Chronic Medications
0543 : Adherence to Statin Therapy for Individuals with Cardiovascular Disease
0545 : Adherence to Statins for Individuals with Diabetes Mellitus
0580 : Bipolar antimanic agent
1879 : Adherence to Antipsychotic Medications for Individuals with Schizophrenia
1927 : Cardiovascular Health Screening for People With Schizophrenia or Bipolar Disorder Who Are Prescribed Antipsychotic Medications
1932 : Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
N/A: Adherence to Antipsychotic Medications for Individuals with Schizophrenia (NCQA measure)

Harmonization

The developer indicates that the measures have been harmonized to the extent possible.

- Measure #1880 is harmonized with related measure #1879 and NCQA version of the measure where possible. The methodology used to calculate adherence, the methodology used to identify the denominator population (with the exception of the clinical conditions), and the data collection burden in all three measures are the same. Three differences exist between the three measures: the clinical codes used to identify the different populations

in each measure; the medications includes in each measure; an exclusion for dementia which is included in NQF Measure #1879 and the NCQA measure but not in measure #1880.

- Measure #1880 has been harmonized to the extent possible with measures #0542, #0543, #0545, #0541, #1879, #1927, and #1932 .
- Measure 1880 has not been harmonized with measure 0580. Measure #0580 differs from measure #1880 because it includes just individuals with newly diagnosed bipolar disorder and major depressive disorder and it identifies the percentage of eligible individuals who have received at least 1 prescription for a mood-stabilizing agent during the measurement year.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 1880

Measure Title: Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document

   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?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐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, skip Questions #3-8, then go to Question #9)

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

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

Signal-to-noise ratio was calculated as a function of the function of variance between physician groups and the variance within a physician group. Reliability was estimated using a beta-binomial model.

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #6)

   ☒Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

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

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☒Yes (go to Question #7)

   ☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2
***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 #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

The developer assessed reliability at the health plan level by calculating inter-rater agreement, using Cohen's Kappa.

8. **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?

REFERENCE: Testing attachment, section 2a2
***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*

☒Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

The final Cohen's Kappa obtained by the two independent programmers were 1.00, which is greater than the Kappa threshold of 0.9.

Part D Plan Reliability (inter-rater agreement)

	Percent Agreement		
Unit of Analysis	**Programmer 1 Num/Den (%)**	**Programmer 2 Num/Den (%)**	**Final Cohen's Kappa**
Part D Plan 1	147/246 (59.8%)	147/246 (59.8%)	1.00
Part D Plan 2	14/32 (43.8%)	14/32 (43.8%)	1.00
Part D Plan 3	52/78 (66.7%)	52/78 (66.7%)	1.00
Part D Plan 4	33/58 (56.9%)	33/58 (56.9%)	1.00
Part D Plan 5	27/50 (54.0%)	27/50 (54.0%)	1.00

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

REFERENCE: testing attachment section 2b1.
NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

***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 #13)

☐No (go to Question #13)

☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

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

REFERENCE: Testing attachment, section 2b4.

☒Yes (please explain below then go to Question #15)

☐No (go to Question #15)

The developer identified a small meaningful difference in measure rates was detected at the plan level with the power of 80% and α=0.05 is 9.6%.

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

REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☐No (go to Question #16)

☒Not applicable (go to Question #16)

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

REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

A small number of missing claims (0.01%) indicates that missing data do not pose a threat to validity.

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; 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 #18)

☒No (please explain below, then skip Questions #18-23 and go to Question #24)

The developer provides justification for why empirical validity testing is not available for this maintenance review and a detailed plan for testing empiric validity before the next maintenance submission on pages 9-11 of the testing attachment.

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

REFERENCE: Testing attachment, section 2b1.

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

☐Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

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

REFERENCE: Testing attachment, section 2b1.

***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 #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☐Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

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

REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)
☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

***REFERENCE:** Testing attachment, section 2b1.*
***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 #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***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 #25)
☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

12 out of a 14 member TEP evaluated the face validity of the measure, indicating strong support of the face validity. 11 of the 12 TEP members "agree" or "strongly agree" that the measure demonstrated face validity. One TEP member who initially voted neutral requested a change to the measure description. Once changed was incorporated changed his/her vote to support the face validity.

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1880
Measure Title: Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder
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: 4/2/2018

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

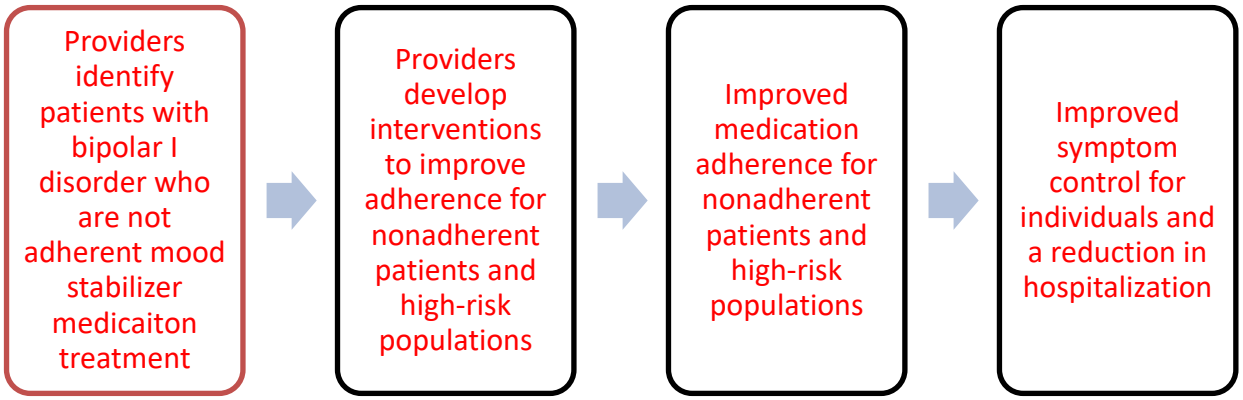
☒ Process: Click here to name what is being measured

☐ 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.



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. This is not a patient-reported measure.

****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)

X 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**	National Institute for Clinical Excellence (NICE)- Bipolar Disorder: Assessment and Management National Institute for Clinical Excellence National Collaborating Centre for Mental Health 2014 The National Institute for Clinical Excellence and the National Collaborating Centre for Mental health. Bipolar Disorder: Assessment and Management. Retrieved from https://www.nice.org.uk/guidance/cg185/evidence/full-guideline-pdf-193212829
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.	If the person is already taking valproate or another mood stabilizers as prophylactic treatment, consider increasing the dose, up to the maximum level in the British National Formulary (BNF) if necessary, depending on clinical response. If there is no improvement, consider adding haloperidol, olanzapine, quetiapine or risperidone, depending on the person's preference and previous response to treatment. If a person develops mania or hypomania and is taking an antidepressant (as defined by the BNF) in combination with a mood stabilizer, consider stopping the antidepressant.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline developers used the GRADE system but did not provide independent grades for each recommendation's evidence. All studies identified evaluating the efficacy of mood stabilizers, lithium and valproate, were rated as having a low quality of evidence.
Provide all other grades and definitions from the evidence grading system	Randomized control trials (RCT) without important limitations provide high quality evidence. Observational studies without special strengths or important limitations provide low quality evidence. For each outcome, quality may be reduced depending on five factors: methodological limitations, inconsistency, indirectness, imprecision and publication bias.
Grade assigned to the **recommendation** with definition of the grade	The Guidelines did not provide independent grades to each recommendation.

Provide all other grades and definitions from the recommendation grading system	The Guidelines did not provide independent grades to each recommendation.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Thirty-six RCTs were included in the body of evidence. The Guideline Development Group found very limited evidence for lithium and valproate monotherapy for acute episodes, but many participants in clinical trials were taking these medications in addition to investigational treatments, and the expert consensus was that mood stabilizers should normally be continued during acute episodes, with doses and plasma levels checked to optimize treatment.
Estimates of benefit and consistency across studies	Most of the studies suffered from very serious limitations, owing to the inappropriate methods that were used for evidence synthesis. According to the remaining studies, valproate semi sodium and lithium (mood stabilizers) were similar in terms of costs and outcomes in an analysis conducted in the US. Olanzapine was found to dominate lithium in a UK study. Quetiapine in addition to mood stabilizer (including quetiapine in XR formulation) was found to be more cost-effective than a mood stabilizer alone in a number of US and UK studies. The existing economic literature review reports conflicting results and is characterized by serious limitations. The guideline cost analysis indicates that lithium may be a cost-effective and potentially cost-saving treatment option for the long-term management of adults with bipolar disorder.
What harms were identified?	Lithium has adverse effects on the kidneys, thyroid and parathyroid. Lithium is a known human teratogen, that is, it is potentially harmful to an unborn child. Valproate is associated with a number of side effects including tremor, weight gain and, rarely, liver damage. It can interact with a number of commonly prescribed medicines and notably is known to decrease plasma levels of olanzapine. Carbamazepine is associated with dizziness, drowsiness, nausea and headaches, and it can cause a low white blood cell count, hyponatremia (low level of sodium in the blood) and rarely, liver damage. Lamotrigine is associated with a rash, drowsiness, dizziness and blurred vision, and it can depress the bone marrow.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Not Applicable

Source of Systematic Review: • **Title**	Practice Guidelines for the Treatment for Patients with Bipolar Disorder

• **Author** • **Date** • **Citation, including page number** • **URL**	American Psychiatric Association 2004 Pyles, R., Cross, C.D., Peele, R., Anzia, D.J., Shemo, J.P., Lurie, L., Walker, R. D., Barnovitz, M.A., Gray, S.H., Saxena, S., and Tonnu, T. (2010). Practice Guidelines for the Treatment of Patients with Bipolar Disorder. American Psychiatric Association. Retrieved from https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/bipolar.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 first-line pharmacological treatment for more severe manic or mixed episodes is the initiation of either lithium plus an antipsychotic or valproate plus an antipsychotic [Recommendation Grade - I]. For less ill patients, monotherapy with lithium, valproate, or an antipsychotic such as olanzapine may be sufficient [Recommendation Grade - I]. For mixed episodes, valproate may be preferred over lithium [Recommendation Grade - II].
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The attributing evidence is not clearly linked to each recommendation, but evidence is linked to specific medications. Each rating of clinical confidence considers the strength of the available evidence and is based on the best available data. When evidence is limited, the level of confidence also incorporates clinical consensus with regard to a particular clinical decision.
Provide all other grades and definitions from the evidence grading system	The following coding system is used to indicate the nature of the supporting evidence in the summary recommendations and references: [A] Double-blind, randomized clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; both the subjects and the investigators are blind to the assignments. [A–] Randomized clinical trial. Same as above but not double-blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally; study does not meet standards for a randomized clinical trial. [C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Case-control study. A study in which a group of patients is identified in the present and information about them is pursued retrospectively or backward in time.

Grade assigned to the **recommendation** with definition of the grade	See brackets after each recommendation above for specific recommendation grades. Overall the grades were: [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence.
Provide all other grades and definitions from the recommendation grading system	The other grade in the recommendation grading system is: [III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Lithium: Five studies have demonstrated lithium is superior to placebo (evidence grade A or B). Three of these studies had randomized assignments, four used crossover designs, and one was a placebo-controlled, parallel design trial. Lithium showed similar efficacy to other mood stabilizers and antipsychotics in 10 other trials (evidence grade A) Valproate: Four randomized placebo-controlled trials (evidence grades A or B) have demonstrated the efficacy of Divalproex/valproate/valproic acid compared to placebo (response rates ranged 48-53%). Valproate was shown to have similar efficacy to other mood stabilizers in four other studies (evidence grades A or B). Olanzapine: Two, large, randomized controlled trials demonstrated that Olanzapine is superior to placebo. Three other randomized controlled trials found similar efficacy to other mood stabilizers (evidence grade A).
Estimates of benefit and consistency across studies	Nearly all studies found that the mood stabilizer was superior for treating bipolar disorder compared to placebo. These studies demonstrated the efficacy of mood stabilizers for every subtype and subgroup of patients with bipolar disorder. Effectiveness of specific medications will vary by patient symptoms and history, see evidence summarized above.
What harms were identified?	Lithium: More common side effects include polyuria, polydipsia, weight gain, cognitive problems, tremor, sedation or lethargy, impaired coordination, gastrointestinal distress, hair loss, benign leukocytosis, acne, and edema. Valproate: More common side effects include sedation, gastrointestinal distress, benign hepatic transaminase elevations, osteoporosis, and tremor. Olanzapine: More common side effects include somnolence, constipation, dry mouth, increased appetite, weight gain, and during titration- orthostatic hypotension.
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.

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.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 1880
Corresponding Measures:
De.2. **Measure Title:** Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services
De.3. **Brief Description of Measure:** Percentage of individuals at least 18 years of age as of the beginning of the measurement period with bipolar I disorder who had at least two prescription drug claims for mood stabilizer medications and had a Proportion of Days Covered (PDC) of at least 0.8 for mood stabilizer medications during the measurement period (12 consecutive months).
1b.1. **Developer Rationale:** We envision several important benefits related to quality improvement with the implementation of this measure. Specifically, the measure will help providers to identify patients with bipolar I disorder who are not adherent (at a critical threshold of 0.8 or greater) with long-term treatment with mood stabilizer medications. Guidelines from the American Psychiatric Association (APA) and the National Institute for Clinical Excellence (NICE) emphasize the importance of treatment adherence and uninterrupted mood stabilizer medication regimens to prevent symptoms and relapse. Furthermore, this measure will encourage providers to develop interventions to improve adherence for this high-risk population. Improved medication adherence among individuals with bipolar I disorder would be expected to result in better control of the initial episode, the prevention of relapse to the initial episode, and the recurrence of new manic or depressive episodes, and as a result, lower mental health-related hospitalization rates and lower suicide rates. APA recommends that pharmacotherapy must be applied in ways that yield good tolerability and do not predispose the patient to nonadherence. Adoption of this performance measure has the potential to improve the quality of care for individuals with bipolar I disorder and, therefore, advance the quality of care in the area of mental health, a priority area identified by the National Priorities Partnership.

S.4. **Numerator Statement:** Individuals with bipolar I disorder who had at least two prescription drug claims for mood stabilizer medications and have a PDC of at least 0.8 for mood stabilizer medications.
S.6. **Denominator Statement:** Individuals at least 18 years of age as of the beginning of the measurement period with bipolar I disorder and at least two prescription drug claims for mood stabilizer medications during the measurement period (12 consecutive months).
S.8. **Denominator Exclusions:** Not Applicable

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Clinician : Group/Practice, Health Plan, Integrated Delivery System, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Mar 04, 2014 **Most Recent Endorsement Date:** Mar 04, 2014

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not Applicable. This measure is not paired.

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall

less-than-optimal performance. ***Measures must be judged to meet all 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
1880_Adherence_to_Mood_Stabilizers_Evidence.docx
1a.1 <u>For Maintenance of Endorsement:</u> 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.
Yes

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)*
<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
We envision several important benefits related to quality improvement with the implementation of this measure. Specifically, the measure will help providers to identify patients with bipolar I disorder who are not adherent (at a critical threshold of 0.8 or greater) with long-term treatment with mood stabilizer medications. Guidelines from the American Psychiatric Association (APA) and the National Institute for Clinical Excellence (NICE) emphasize the importance of treatment adherence and uninterrupted mood stabilizer medication regimens to prevent symptoms and relapse. Furthermore, this measure will encourage providers to develop interventions to improve adherence for this high-risk population. Improved medication adherence among individuals with bipolar I disorder would be expected to result in better control of the initial episode, the prevention of relapse to the initial episode, and the recurrence of new manic or depressive episodes, and as a result, lower mental health-related hospitalization rates and lower suicide rates. APA recommends that pharmacotherapy must be applied in ways that yield good tolerability and do not predispose the patient to nonadherence. Adoption of this performance measure has the potential to improve the quality of care for individuals with bipolar I disorder and, therefore, advance the quality of care in the area of mental health, a priority area identified by the National Priorities Partnership.

1b.2. **Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis**. *(<u>This is required for maintenance of endorsement</u>. 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.*
TESTING RESULTS BASED ON MEDICARE DATA
FMQAI (now HSAG) analyzed Medicare administrative data from eight states and calculated measure rates as part of the testing of this measure. Although our results suggest better adherence in the Medicare population than some published studies (described below), we still identified substantial performance gaps and wide variation in adherence to mood stabilizer medications with a PDC of 0.8 or greater among persons with bipolar I disorder across states, Part D Plans, Accountable Care Organizations (ACOs), and physician groups. The overall measure rate across eight states was 67.2%, indicating that 1 of 3 individuals with bipolar I disorder taking mood stabilizer medications has an adherence rate less than 0.8. The measure rates for the eight states ranged from 60.8% to 77.4%, and the rates among plans with at least 30 individuals in the denominator ranged from 53.4% to 77.1%, ACOs with at least 30 individuals in the denominator ranged from 51.0% to 77.0%, and physician groups with at least 30 individuals in the denominator had more variability than the other units analyzed, ranging from 44.3% to 90.5%.

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.**
Eight studies (Bagalman et al., 2010; Berger et al., 2012; Hajda et al., 2015; Hong et al., 2011; Lage et al., 2009; Lang et al., 2011; Lew et al., 2006; Rascati et al., 2011) demonstrate low rates of adherence among individuals with bipolar I disorder who are prescribed mood stabilizer medications. These low adherence rates were corroborated by the results of measure testing conducted by FMQAI (now HSAG) of Medicare data, which also showed considerable variation among providers. Both the low rates of adherence and variation among providers indicate a performance gap in the treatment of individuals with bipolar I disorder. Reported rates of adherence to mood stabilizer medications (defined as a PDC or MPR of 0.8 or greater) among persons with bipolar I disorder range from 16% to 76% in these studies. The published studies and the testing results are described below.

SUMMARY OF PUBLISHED STUDIES ON VARIATION IN PERFORMANCE

BAGALMAN ET AL. (2010): This study used 2000-2005 claims data for 1,258 commercially insured persons with bipolar disorder to estimate adherence. About one third (35.7%) were classified as adherent (MPR of at least 0.8), based on the 12 months following an index prescription.

BERGER ET AL. (2012): This study was a retrospective cohort analysis of administrative data on 84 patients with bipolar disorder hospitalized between 2001 and 2008 (mean age of 45 years) (Berger et al., 2012). During the six months following the hospitalization for bipolar disorder, only 15.5% of these patients had an MPR of over 80% for the antipsychotic medication initially prescribed at the time of discharge. An additional 26% had switched to another antipsychotic agent by 6 months.

HAJDA ET AL. (2015): This study was a cross sectional study of 33 outpatients with bipolar disorder who completed a scale to estimate treatment adherence. The study found that more than half (57.6%) of the patients with bipolar disorder had discontinued medication previously. The risk of the discontinuation of medication was higher in patients who were young and single. The rate of current adherence was significantly negatively correlated with self-stigma.

HONG ET AL. (2011): This study was a prospective observational study that followed 1,341 patients (18 years and older) with bipolar disorder for 21 months after a manic/mixed episode in 2002-2004. In this study, 76.4% of patients were classified as adherent to a bipolar disorder medication (antipsychotics, anticonvulsants, and/or lithium), based on a psychiatrist´s assessment.

LAGE ET AL. (2009): This study was a retrospective analysis of claims data for commercial health plans on 7,769 patients with bipolar disorder who were 18-64 years of age. In this study, the mean MPR for antipsychotics was 41.7%, with 61.9% of patients having an MPR =0.50 and 78.7% having an MPR =0.75.

LANG ET AL. (2011): This study was a retrospective cohort analysis of 2004-2007 claims for 9,410 Medicaid patients with bipolar I disorder (mean age of 38 years). In this study, 60% of Medicaid patients were nonadherent (MPR less than 0.8) to antipsychotic medications during the year following their first antipsychotic prescription based on claims data.

LEW ET AL. (2006): This study was a retrospective analysis of prescription and medical claims for a large managed care organization representing commercial health plan members. An estimated 45.2% of 1,399 patients had an adherence rate of at least 0.80 to traditional mood-stabilizing therapy (lithium, valproate, carbamazepine, lamotrigine, or oxcarbazepine).

RASCATI ET AL. (2011): This study analyzed 2002-2008 Medicaid claims data for 2,446 Medicaid patients with bipolar disorder to assess adherence rates for second-generation antipsychotic (SGA) medications. Of those receiving a prescription, 58% were adherent (MPR of at least 0.8) during the 12 months following the first prescription.

CONCLUSION

Estimates of adherence to mood stabilizer medications among individuals with bipolar I disorder from recently published studies and our testing results suggest a clear performance gap. For reference, the published studies reported the adherence rates to mood stabilizer medications (defined as PDC or MPR of 0.8 or greater), ranging from 16% to 76%. The measure rate for the eight states based on Medicare data ranged from 60.8% to 77.4%. These rates represent performance gaps, variation, and opportunities for improvement in the treatment of individuals with bipolar I disorder.

References:

Bagalman, E., Yu-Isenberg, K. S., Durden, E., Crivera, C., Dirani, R., and Bunn, W. B. 3rd. (2010). Indirect costs associated with nonadherence to treatment for bipolar disorder. J Occup Environ Med, 52(5), 478-85.

Berger, A., Edelsberg, J., et al. (2012). Medication adherence and utilization in patients with schizophrenia or bipolar disorder receiving aripiprazole, quetiapine, or xiprasidone at hospital discharge: A retrospective cohort study. BMC Psychiatry, 12, 99.

Hajda, M., Kamaradova, D., Latalova, K., Prasko, J., Ociskova, M., Mainerova, B., Cinculova, A., Vrbova, K., Kubinek, R., and Tichackova, A. Self-stigma, treatment adherence, and medication discontination in patients with bipolar disorders in remission - cross sectional study. Activitas Nervosa Superior Rediviva, 57 (1-2), 6-11. Ebpub 2015 Apr 1.

Hong, J., Reed, C., Novick, D., Haro, J. M., and Aguado, J. (2011). Clinical and economic consequences of medication non-adherence in the treatment of patients with a manic/mixed episode of bipolar disorder: Results from the European Mania in Bipolar Longitudinal Evaluation of Medication (EMBLEM) study. Psychiatry Res, 190(1), 110-4. Epub 2011 May 14.

Lage, M. and Hassan, M. (2009). The relationship between antipsychotic medication adherence and patient outcomes among individuals diagnosed with bipolar disorder: a retrospective study. Ann Gen Psychiatry, 8, 7.

Lang, K., Korn, J., Muser, E., Choi, J. C., Abouzaid, S., and Menzin, J. (2011). Predictors of medication nonadherence and hospitalization in Medicaid patients with bipolar I disorder given long-acting or oral antipsychotics. J Med Econ, 14(2), 217-26. Epub 2011 Mar 4.

Lew, K. H., Chang, E. Y., et al. (2006). The effect of medication adherence on health care utilization in bipolar disorder. Managed Care Interface, 19(9), 41-46.

Rascati, K., Richards, K., et al. (2011). Adherence, persistence of use, and costs associated with second-generation antipsychotics for bipolar disorder. Psychiatric Services, 62(9), 1032-1040.

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.*

TESTING RESULTS BASED ON MEDICARE DATA

We analyzed 2007-2008 claims data for 27,798 Medicare beneficiaries with bipolar I disorder. A consistent pattern was observed with adherence rates for mood stabilizer medications being substantially lower among African-American and Hispanic persons with bipolar I disorder compared with White persons. For all age groups combined, the adherence rates (i.e., proportion of days covered of at least 0.8) for all ages were 55.3% and 62.6% for African-American and Hispanic persons, respectively, and 68.6% for White persons. The adherence rates were lower among African-American and Hispanic persons than among White persons in every age group, except 65-74 and 85 and older, in which African-American rates were higher than White rates. However, African-American rates were lower than Hispanic rates in some age groups (i.e., 25-44, 45-64, and 75-84 years), and higher in all other age groups (i.e., 18-24, 65-74, and 85+ years).

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

SUMMARY OF PUBLISHED STUDIES ON DISPARITIES BY POPULATION GROUP

The four studies described in this section (Garcia, et al., 2016; Rascati et al., 2011; Sajatovic et al., 2006; Zeber et al., 2011) reported higher adherence rates among White persons with bipolar I disorder than among African-American and Hispanic persons with bipolar I disorder. One recent study also found age and education to be associated with adherence rates.

GARCIA ET AL. (2016): This systematic review found age, race, and education to be associated with adherence rates. Younger patients were less adherent than older patients, African-American patients had lower adherence rates than White patient, and patients with lower levels of education had poorer adherence. The review found economic and transportation barriers hinder patient's adherence to treatment.

RASCATI ET AL. (2011): This study assessed adherence rates to second-generation antipsychotic (SGA) medications among 2,446 Medicaid patients with bipolar disorder based on 2002-2008 Medicaid claims data. African-American and Hispanic patients were more likely than White patients to have poor adherence (MPR less than 0.8) to second-generation antipsychotic medication during the 12 months following the first prescription (odds ratio=1.97 and 1.35, respectively).

SAJATOVIC ET AL. (2006): Based on a retrospective analysis of adherence data on 26,986 veterans with a bipolar disorder diagnosis who were prescribed an antipsychotic medication during fiscal year 2003, Sajatovic et al. (2006) reported counts of patients by adherence and ethnicity. Based on these data, Whites had higher adherence rates than African-Americans and Hispanics: 55%, 38%, and 50% of Whites, African-Americans, and Hispanics, respectively, were fully adherent (MPR of at least 0.8)

with antipsychotic medication; 21%, 25%, and 22%, respectively, were partially adherent (MPR of at least 0.5 and less than 0.8); and 24%, 37%, and 28%, respectively, were non-adherent (MPR less than 0.5).

ZEBER ET AL. (2011): In a cross-sectional population-based study of 435 VA patients with bipolar disorder, poor adherence was found to be self-reported more often by ethnic minorities (i.e., primarily African-Americans) (60%) than White veterans (42%). In addition, a higher percentage of two minority groups reported missing some recent medication doses (39%), compared to 23% of White patients (p <0.01 on both adherence measures).

CONCLUSION
In regard to age-related disparities, adherence rates were lower among persons 18-64 years of age than among those 65 years of age and over. This pattern of lower adherence rates in younger persons was consistent for White and African-American persons and for all age groups except a higher rate among Hispanic persons 45-64 years of age.

References:
Garcia, S., Martínez-Cengotitabengoa, M., López-Zurbano, S., et al. (2016). Adherence to antipsychotic medication in bipolar disorder and schizophrenic patients: A systematic review. Journal of Clinical Psychopharmacology, 36(4), 355-371.

Rascati, K., Richards, K., et al. (2011). Adherence, persistence of use, and costs associated with second-generation antipsychotics for bipolar disorder. Psychiatric Services, 62(9), 1032-1040.

Sajatovic, M., Valenstein, M., Blow, F. C., Ganoczy, D., and Ignacio, R. V. (2006). Treatment adherence with antipsychotic medications in bipolar disorder. Bipolar Disord, 8, 232-241.

Zeber, J. E., Miller, A. L., Copeland, L. A., McCarthy, J. F., Zivin, K., Valenstein, M., et al. (2011). Medication adherence, ethnicity, and the influence of multiple psychosocial and financial barriers. Adm Policy Mental Health, 38(2), 86-95.

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):*
Disparities Sensitive

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Elderly, Populations at Risk, Populations at Risk : Dual eligible beneficiaries

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 Applicable

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:** NQF_1880_Code_Tables_2018_Final.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.
Yes

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.
• Updated NDCs as of March 9, 2018
• Added medications with FDA approval for the treatment of bipolar I disorder: cariprazine, quetiapine fumarate (Seroquel)
• Removed medications lacking FDA approval for the treatment of bipolar I disorder: fluphenazine, haloperidol, molindone, perphenazine, pimozide, prochlorperazine, thioridazine, thiothixene, trifluoperazine, clozapine, iloperidone, paliperidone, fluphenazine decanoate, haloperidol decanoate, olanzapine pamoate, paliperidone palmitate
• Added the following code to the value set for identifying bipolar I disorder: F30.8 (other manic episodes)

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).
Individuals with bipolar I disorder who had at least two prescription drug claims for mood stabilizer medications and have a PDC of at least 0.8 for mood stabilizer medications.

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 defined as individuals with a PDC of 0.8 or greater.

The PDC is calculated as follows:
PDC NUMERATOR
The PDC numerator is the sum of the days covered by the days' supply of all prescription drug claims for all mood stabilizer medications. The period covered by the PDC starts on the day the first prescription is filled (index date) and lasts through the end of the measurement period, or death, whichever comes first. For prescriptions drug claims with a days' supply that extends beyond the end of the measurement period, count only the days for which the drug was available to the individual during the measurement period. If there are claims for the same drug (generic name) on the same date of service, keep the claim with the largest days' supply. If claims for the same drug (generic name) overlap, then adjust the prescription start date to be the day after the previous fill has ended.

PDC DENOMINATOR
The PDC denominator is the number of days from the first prescription drug claim date through the end of the measurement period, or death date, whichever comes first.

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

Individuals at least 18 years of age as of the beginning of the measurement period with bipolar I disorder and at least two prescription drug claims for mood stabilizer medications during the measurement period (12 consecutive months).

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. Continuously enrolled in Medicare Part D with no more than a one-month gap in enrollment during the measurement year;
2. Continuously enrolled in Medicare Part A and Part B with no more than a one-month gap in Part A enrollment and no more than a one-month gap in Part B enrollment during the measurement year; and,
3. No more than one month of HMO (Health Maintenance Organization) enrollment during the measurement year.

IDENTIFICATION OF BIPOLAR I DISORDER
Individuals with bipolar I disorder are identified by having a diagnosis of bipolar I disorder within the inpatient or outpatient claims data. Individuals must have:

At least two encounters with a diagnosis of bipolar I disorder with different dates of service in an outpatient setting, emergency department setting, or non-acute inpatient setting during the measurement period;

OR

At least one encounter with a diagnosis of bipolar I disorder in an acute inpatient setting during the measurement period.

CODES USED TO IDENTIFY BIPOLAR I DISORDER DIAGNOSIS
Codes used to identify bipolar I disorder are included in the attached Excel worksheet of codes (NQF_1880_Code Tables_2018 Final) under the tab NQF_1880_Bipolar_ICD9-10.
TABLE 1. BIPOLAR I DISORDER DIAGNOSIS
ICD-9-CM: 296.0x, 296.1x, 296.4x, 296.5x, 296.6x, 296.7
ICD-10-CM: F30.10, F30.11, F30.12, F30.13, F30.2, F30.3, F30.4, F30.8, F30.9, F31.0, F31.10, F31.11, F31.12, F31.13, F31.2, F31.30, F31.31, F31.32, F31.4, F31.5, F31.60, F31.61, F31.62, F31.63, F31.64, F31.70, F31.71, F31.72, F31.73, F31.74, F31.75, F31.76, F31.77, F31.78, F31.89, F31.9

CODES USED TO IDENTIFY ENCOUNTER TYPE
Codes used to identify encounters are under tab NQF_1880_Encounter_types.

TABLE 2.1. OUTPATIENT SETTING
Current Procedural Terminology (CPT): 98960-98962, 99078, 99201-99205, 99211-99215, 99217-99220, 99241-99245, 99341-99345, 99347-99350, 99385-99387, 99395-99397, 99401-99404, 99411, 99412, 99429, 99510
HCPCS: G0155, G0176, G0177, G0409-G0411, G0463, H0002, H0004, H0031, H0034-H0037, H0039, H0040, H2000, H2001, H2010-H2020, M0064, S0201, S9480, S9484, S9485, T1015
UB-92 revenue: 0510, 0511, 0513, 0516-0517, 0519-0523, 0526-0529, 0770, 0771, 0779, 0900-0905, 0907, 0911-0917, 0919, 0982, 0983

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 90880, 99221-99223, 99231-99233, 99238, 99239, 99251-99255, 99291

WITH

Place of Service (POS): 03, 05, 07, 09, 11, 12, 13, 14, 15, 20, 22, 24, 33, 49, 50, 52, 53, 71, 72

TABLE 2.2. EMERGENCY DEPARTMENT SETTING
CPT: 99281-99285

UB-92 revenue: 0450, 0451, 0452, 0456, 0459, 0981

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 99291

WITH

POS: 23

TABLE 2.3. NON-ACUTE INPATIENT SETTING
CPT: 99304-99310, 99315, 99316, 99318, 99324-99328, 99334-99337
HCPCS: H0017-H0019, T2048
UB-92 revenue: 0118, 0128, 0138, 0148, 0158, 0190-0194, 0199, 0524, 0525, 0550-0552, 0559, 0660-0663, 0669, 1000, 1001, 1003-1005

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 99291

WITH

POS: 31, 32, 56

TABLE 2.4. ACUTE INPATIENT SETTING
UB-92 revenue: 0100, 0101, 0110-0114, 0119-0124, 0129-0134, 0139-0144, 0149-0154, 0159, 0160, 0164, 0167, 0169, 0200-0204, 0206-0209, 0210-0214, 0219, 0720-0724, 0729, 0987

OR

CPT: 90791, 90792, 90832-90834, 90836-90840, 90845, 90847, 90849, 90853, 90863, 90867-90870, 90875, 90876, 99221-99223, 99231-99233, 99238, 99239, 99251-99255, 99291

WITH

POS: 21, 51

IDENTIFICATION OF PRESCRIPTION DRUG CLAIMS FOR MOOD STABILIZER MEDICATION

Individuals with at least two prescription drug claims for any of the following mood stabilizer medications (Table 3: Mood Stabilizer Medications) or long-acting injectable antipsychotic medications (see Table 4: Long-acting injectable antipsychotic medications). The National Drug Center (NDC) identifier for medications included in the measure denominator are listed in tab NQF_1880_Mood_Stabilizers of the attached Excel workbook. Obsolete drug products are excluded from National Drug Codes (NDCs) with an inactive date more than six years prior to the beginning of the measurement period or look-back period.

MOOD STABILIZER MEDICATIONS

TABLE 3. MOOD STABILIZER MEDICATIONS
Active ingredients listed below are limited to oral, buccal, sublingual, and translingual formulations only.

Anticonvulsants:
carbamazepine
divalproex sodium
lamotrigine
valproic acid

Atypical Antipsychotics:
aripiprazole
asenapine
cariprazine
lurasidone
olanzapine
quetiapine
quetiapine fumarate (Seroquel)
risperidone
ziprasidone

Phenothiazine/Related Antipsychotics:
chlorpromazine
loxapine succinate

Other Antipsychotics:
olanzapine-fluoxetine

Lithium Salts:
lithium carbonate
lithium citrate

TABLE 4: LONG-ACTING INJECTABLE ANTIPSYCHOTIC MEDICATIONS
The following are the long-acting (depot) injectable antipsychotic medications. The route of administration includes all injectable and intramuscular formulations of the medications listed below.

Atypical Antipsychotic Medications:
aripiprazole (J0401)
risperidone microspheres (J2794)

Note: Since the days' supply variable is not reliable for long-acting injections in administrative data, the days' supply is imputed as listed below for the long-acting (depot) injectable antipsychotic medications billed under Medicare Part D and Part B:
aripiprazole (J0401) – 28 days' supply
risperidone microspheres (J2794) – 14 days' supply

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Not Applicable

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.)*
Depending on the operational use of the measure, measure results may be stratified by:
- State
- Accountable Care Organization (ACOs)*
- Plan
- Physician Group**
- Age – Divided into six categories: 18-24, 25-44, 45-64, 65-74, 75-84, and 85+ years
- Race/Ethnicity

• Dual Eligibility

*ACO attribution methodology is based on where the beneficiary is receiving the plurality of his/her primary care services and subsequently assigned to the participating providers.
**See Calculation Algorithm/Measure Logic S.14 below for physician group attribution methodology used for this measure.

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.)*
Target Population: Individuals at least 18 years of age as of the beginning of the measurement period who have met the enrollment criteria for Medicare Parts A, B, and D.

Denominator: Individuals at least 18 years of age as of the beginning of the measurement period with bipolar I disorder and at least two prescription drug claims for mood stabilizer medications during the measurement period (12 consecutive months).

CREATE DENOMINATOR:
1. Pull individuals who are 18 years of age or older as of the beginning of the measurement period.
2. Include individuals who were continuously enrolled in Medicare Part D coverage during the measurement period, with no more than a one-month gap in enrollment during the measurement period, or up until their death date if they died during the measurement period.
3. Include individuals who had no more than a one-month gap in Medicare Part A enrollment, no more than a one-month gap in Part B enrollment, and no more than one month of HMO (Health Maintenance Organization) enrollment during the current measurement period (fee-for-service [FFS] individuals only).
4. Of those individuals identified in Step 3, keep those who had:
At least two encounters with a diagnosis of bipolar I disorder with different dates of service in an outpatient setting, emergency department setting, or non-acute inpatient setting during the measurement period;
OR
At least one encounter with a diagnosis of bipolar I disorder in an acute inpatient setting during the measurement period.
5. Of the individuals identified in Step 4, extract Medicare Part D claims for a mood stabilizer during the measurement period. Attach the drug ID and the generic name to the dataset.
6. For the individuals identified in Step 5, exclude those who did not have at least two prescription drug claims for any mood stabilizer on different dates of service (identified by having at least two Medicare Part D claims with the specific codes) during the measurement period.

Numerator: Individuals with bipolar I disorder who had at least two prescription drug claims for mood stabilizer medications and have a PDC of at least 0.8 for mood stabilizer medications.

CREATE NUMERATOR:
For the individuals in the denominator, calculate the PDC for each individual according to the following methods:
1. Determine the individual's medication therapy period, defined as the index prescription date through the end of the measurement period, or death, whichever comes first. The index date is the service date (fill date) of the first prescription drug claim for a mood stabilizer medication in the measurement period.
2. Within the medication therapy period, count the days the individual was covered by at least one drug in the mood stabilizer medication class based on the prescription drug claim service date and days of supply.

a. Sort and de-duplicate Medicare Part D claims for mood stabilizers by beneficiary ID, service date, generic name, and descending days' supply. If prescriptions for the same drug (generic name) are dispensed on the same date of service for an individual, keep the dispensing with the largest days' supply.
b. Calculate the number of days covered by mood stabilizer therapy per individual.
i. For prescription drug claims with a days' supply that extends beyond the end of the measurement period, count only the days for which the drug was available to the individual during the measurement period.
ii. If claims for the same drug (generic name) overlap, then adjust the latest prescription start date to be the day after the previous fill has ended.
iii. If claims for different drugs (different generic names) overlap, do not adjust the prescription start date.
3. Calculate the PDC for each individual. Divide the number of covered days found in Step 2 by the number of days in the individual's medication therapy period found in Step 1.

An example of SAS code for Steps 1-3 was adapted from Pharmacy Quality Alliance (PQA) and is also available at the URL: http://www2.sas.com/proceedings/forum2007/043-2007.pdf.

4. Of the individuals identified in Step 3, count the number of individuals with a calculated PDC of at least 0.8 for the mood stabilizers. This is the numerator.

PHYSICIAN GROUP ATTRIBUTION:
Physician group attribution was adapted from Generating Medicare Physician Quality Performance Measurement Results (GEM) Project: Physician and Other Provider Grouping and Patient Attribution Methodologies (http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/GEM/downloads/GEMMethodologies.pdf). The following is intended as guidance and reflects only one of many methodologies for assigning individuals to a medical group. Please note that the physician group attribution methodology excludes patients who died, even though the overall measure does not.

I. Identify Physician and Medical Groups
1. Identify all Tax Identification Numbers (TINs)/National Provider Identification (NPI) combinations from all Medicare Part B claims in the measurement year and the prior year. Keep records with valid NPIs. Valid NPIs have 10 numeric characters (no alpha characters).
2. For valid NPIs, pull credentials and specialty code(s) from the CMS provider tables.
3. Create one record per NPI with all credentials and all specialties. A provider may have more than one specialty.
4. Attach TIN to NPI, keeping only those records with credentials indicating a physician (MD or DO), physician assistant (PA), or nurse practitioner (NP).
5. Identify medical group TINs: Medical group TINs are defined as TINs that had physician, physician assistant, or nurse practitioner provider specialty codes on at least 50% of Medicare Part B carrier claim line items billed by the TIN during the measurement year or prior year. (The provider specialty codes are listed after Patient Attribution.)
a. Pull Part B records billed by TINS identified in Step 4 during the measurement year and prior year.
b. Identify claims that had the performing NPI (npi_prfrmg) in the list of eligible physicians/TINs, keeping those that match by TIN, performing NPI, and provider state code.
c. Calculate the percentage of Part B claims that match by TIN, npi_prfrmg, and provider state code for each TIN, keeping those TINs with percentages greater than or equal to 50%.
d. Delete invalid TINs. Examples of invalid TINs are defined as having the same value for all nine digits or values of 012345678, 012345678, 123456789, 987654321, or 87654321.
6. Identify TINs that are not solo practices.
a. Pull Part B records billed by physicians identified in Step 4 for the measurement year and/or prior year.
b. Count unique NPIs per TIN.
c. Keep only those TINs having two or more providers.
d. Delete invalid TINs. Examples of invalid TINs are defined as having the same value for all nine digits or values of 012345678, 012345678, 123456789, 987654321, or 87654321.
7. Create final group of TINs from Step 5 and Step 6 (TINs that are medical groups and are not solo practices).
8. Create file of TINs and NPIs associated with those TINs. These are now referred to as the medical group TINs.
9. Determine the specialty of the medical group (TIN) to be used in determining the specialty of nurse practitioners and physician assistants. The plurality of physician providers in the medical group determines the specialty of care for nurse practitioners and physician assistants.
a. From the TIN/NPI list created in Step 8, count the NPIs per TIN/specialty.
b. The specialty with the maximum count is assigned to the medical group.

II. Identify Individual Sample and Claims
10. Create individual sample.
a. Pull individuals with 11+ months of Medicare Parts A, B, and D during the measurement year.
b. Verify the individual did not have any months with Medicare as secondary payer. Remove individuals with BENE_PRMRY_PYR_CD not equal to one of the following:
• A = working-age individual/spouse with an employer group health plan (EGHP)
• B = End Stage Renal Disease (ESRD) in the 18-month coordination period with an EGHP
• G = working disabled for any month of the year
c. Verify the individual resides in the U.S., Puerto Rico, Virgin Islands, or Washington D.C.
d. Exclude individuals who enter the Medicare hospice at any point during the measurement year.
e. Exclude individuals who died during the measurement year.
11. For individuals identified in Step 10, pull office visit claims that occurred during the measurement year and in the six months prior to the measurement year.
a. Office visit claims have CPT codes of 99201-99205, 99211-99215, and 99241-99245.
b. Exclude claims with no npi_prfrmg.
12. Attach medical group TIN to claims by NPI.

III. Patient Attribution
13. Pull all Medicare Part B office claims from Step 12 with specialties indicating primary care or psychiatry (see list of provider specialties and specialty codes below). Attribute each individual to at most one medical group TIN for each measure.
a. Evaluate specialty on claim (HSE_B_HCFA_PRVDR_SPCLTY_CD) first. If specialty on claim does not match any of the measure-specific specialties, then check additional specialty fields.
b. If the provider specialty indicates nurse practitioners or physician assistants (code 50 or code 97), then assign the medical group specialty determined in Step 9.
14. For each individual, count claims per medical group TIN. Keep only individuals with two or more E&M claims.
15. Attribute the individual to the medical group TIN with the most claims. If a tie occurs between medical group TINs, attribute the TIN with the most recent claim.
16. Attach the medical group TIN to the denominator and numerator files by individual.

Provider Specialties and Specialty Codes
Provider specialties and specialty codes include only physicians, physician assistants, and nurse practitioners for physician grouping, TIN selection, and patient attribution. The provider specialty codes and the associated provider specialty are shown below:

01—General practice*
02—General surgery
03—Allergy/immunology
04—Otolaryngology
05—Anesthesiology
06—Cardiology
07—Dermatology
08—Family practice*
09—Interventional pain management
10—Gastroenterology
11—Internal medicine*
12—Osteopathic manipulative therapy
13—Neurology
14—Neurosurgery
16—Obstetrics/gynecology*
18—Ophthalmology
20—Orthopedic surgery
22—Pathology
24—Plastic and reconstructive surgery
25—Physical medicine and rehabilitation
26—Psychiatry*
28—Colorectal surgery
29—Pulmonary disease

30—Diagnostic radiology
33—Thoracic surgery
34—Urology
36—Nuclear medicine
37—Pediatric medicine
38—Geriatric medicine*
39—Nephrology
40—Hand surgery
44—Infectious disease
46—Endocrinology
50—Nurse practitioner*
66—Rheumatology
70—Multi-specialty clinic or group practice*
72—Pain management
76—Peripheral vascular disease
77—Vascular surgery
78—Cardiac surgery
79—Addiction medicine
81—Critical care (intensivists)
82—Hematology
83—Hematology/oncology
84—Preventive medicine*
85—Maxillofacial surgery
86—Neuropsychiatry*
90—Medical oncology
91—Surgical oncology
92—Radiation oncology
93—Emergency medicine
94—Interventional radiology
97—Physician assistant*
98—Gynecologist/oncologist
99—Unknown physician specialty
Other—NA
*Provider specialty codes specific to this measure

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 does not use a sample or survey.

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.

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.
For measure calculation in the Medicare product line, the following Medicare files were required:
• Denominator tables
• Prescription drug benefit (Part D) coverage tables
• Beneficiary file
• Institutional claims (Part A)
• Non-institutional claims (Part B)—physician carrier/non-DME

• Prescription drug benefit (Part D) claims

For ACO attribution, the following were required:
• Denominator tables for Parts A and B enrollment
• Prescription drug benefit (Part D) coverage tables
• Beneficiary file
• Institutional claims (Part A)
• Non-institutional claims (Part B)—physician carrier/non-DME
• Prescription drug benefit (Part D) claims

For physician group attribution, the following were required:
• Non-institutional claims (Part B)—physician carrier/non-DME
• Denominator tables to determine individual enrollment
• Beneficiary file or coverage table to determine hospice benefit and Medicare as secondary payor status
• CMS physician and physician specialty tables
• National Plan and Provider Enumeration System (NPPES) database

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, Health Plan, Integrated Delivery System, 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.)*

2. Validity – See attached Measure Testing Submission Form
1880_Adherence_to_Mood_Stabilizers_Testing-636582869208053114.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.
Yes

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.
No

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 1880
Measure Title: Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder
Date of Submission: 4/2/2018
Type of Measure:

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

Instructions

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

Medicare Parts A, B, and D claims data and Minimum Data Set (MDS) data for calendar years 2007 and 2008 were used to support the field testing of the measure.

Additional data used in testing included Parts A, B, and D data for beneficiaries in 32 ACOs from calendar year 2010.

For measure calculation, the following Medicare files were required:

- Denominator tables
- Prescription drug benefit (Part D) coverage tables
- Beneficiary file
- Institutional claims (Part A)
- Non-institutional claims (Part B)—physician carrier/non-DME
- Prescription drug benefit (Part D) claims

For ACO attribution, the following were required:

- Denominator tables for Parts A and B enrollment
- Prescription drug benefit (Part D) coverage tables
- Beneficiary file
- Institutional claims (Part A)
- Non-institutional claims (Part B)—physician carrier/non-DME
- Prescription drug benefit (Part D) claims

For physician group attribution, the following were required:

- Non-institutional claims (Part B)—physician carrier/non-DME
- Denominator tables to determine individual enrollment
- Beneficiary file or coverage table to determine hospice benefit and Medicare as secondary payor status
- CMS physician and physician specialty tables
- National Plan & Provider Enumeration System (NPPES) database

1.3. What are the dates of the data used in testing? 2007, 2008, 2010

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: integrated delivery system, population (state)	☒ other: integrated delivery system, population (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 from eight states (Arizona, Delaware, Florida, Iowa, Indiana, Mississippi, Rhode Island, and Washington) were included in the testing and analysis for both reliability and validity. These data included 9,406 Physician Groups and 656 Part D plans.

Additional data used in testing included Parts A, B, and D data for beneficiaries in 32 ACOs from calendar year 2010.

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

The data included 4,789,034 Medicare beneficiaries.

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.

No differences in the data or sample used.

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.

Two proxy variables for social risk were evaluated to understand disparities: race/ethnicity and dual-eligibility beneficiary status. Because this measure is not an outcome or intermediate outcome measure, these factors were not evaluated for risk adjustment. Overall, African-Americans and non-dually eligible individuals under age 85 had rates about 10%-15% lower than other groups.

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*)

In order to assess measure precision in the context of the observed variability across measurement units (states, prescription drug plans [serving as a proxy for health plans], Accountable Care Organizations [ACOs], and physician groups), we utilized the approach proposed by Adams (2009) and Scholle et al. (2008). The rationale for this choice of testing was based on the work on the reliability for provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial published by Adams: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient."

The signal-to-noise ratio was calculated as a function of the variance between physician groups (signal) and the variance within a physician group (noise). Reliability was estimated using a beta-binomial model. This approach has 2 basic assumptions:

1) Each physician has a true pass rate, p, which varies from physician to physician, and
2) The physician's score is a binomial random variable conditional on the physician's true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual physician group variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across physician groups). In a simulation, Adams showed that differences between physicians started to be seen at reliability of 0.7 and significant differences could be seen at reliability of 0.9. Our rationale was based on Adams' work, and thus, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to discriminate performance between physicians.

Using methodology described by Scholle et al. (2008), reliability estimates were computed separately based on the mean denominator size for physicians within each denominator category. As Scholle described in the article, the reliability estimate at the mean denominator for each category should reflect "the typical experience of physicians in this population."

Reliability scores were also calculated for states, Part D plans (which served as a proxy for health plans), and Accountable Care Organizations using the same approach.

Reliability at the health plan level was also assessed using Cohen's Kappa. The measure scores for five randomly selected Medicare Part D plans from two states (Florida and Rhode Island) were compared, and inter-rater agreement was calculated. Concerning an acceptable threshold for kappa, there are no definitive criteria in the literature for what level of reliability is acceptable for measures based on administrative data. Furthermore, since relatively small differences in programmer interpretation could result in a large variation in output, we utilized a conservative threshold of 0.9 for Cohen's Kappa, based on the following scale:

< 0 = no agreement
0–0.20 = slight agreement
0.21–0.40 = fair agreement
0.41–0.60 = moderate agreement
0.61–0.80 = substantial agreement
0.81–1 = almost perfect agreement

Therefore, if the Cohen's Kappa was greater than or equal to 0.9, the measure specifications were considered reliable. If Cohen's Kappa in the initial reliability testing with the two programmers was less than 0.9, each step of the measure algorithm (in the Measure Information Form [MIF]) was compared, and the differences were clarified between programmer 1 and 2. Identified differences are noted in a narrative, where applicable, along with extracts of the respective modification to the MIF.

The revised MIF was then presented to a third programmer and results compared to the consolidated results derived in the first round of reliability testing. This iterative process with independent programmers continued until the Kappa score reached the threshold of greater than or equal to 0.9.

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

Scholle, S. H., Roski, J., Adams, J. L., Dunn, D. L., Kerr, E. A., Dugan, D. P., et al. (2008). Benchmarking physician performance: Reliability of individual and composite measures. *American Journal of Managed Care*, 14(12), 833-838

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*)

State Reliability
State / Denominator / Mean rate for state / Reliability score (based on the mean rate)

DE / 679 / 63.77% / 0.896

RI / 931 / 69.60% / 0.928

AZ / 1,376 / 60.76% / 0.944

MS / 2,187 / 61.00% / 0.964

IA / 2,292 / 77.40% / 0.974

WA / 3,649 / 71.77% / 0.981

IN / 4,781 / 70.09% / 0.985

FL / 11,962 / 64.61% / 0.994

Part D Plan Reliability (signal-to-noise analysis)
Minimum denominator size of Part D plan / # of Plans / Mean denominator size of Part D plan / Mean rate of Part D plans / Reliability score (based on the mean rate and the mean denominator size) / Reliability score (based on the mean rate and the minimum denominator)

10 / 34 / 817 / 66.51 / 0.6070 / 0.0454
20 / 29 / 956 / 66.53 / 0.6438 / 0.0847
30 / 27 / 1,025 / 66.49 / 0.6814 / 0.1236
50 / 26 / 1,063 / 66.08 / 0.6893 / 0.1892
100 / 22 / 1,245 / 65.88 / 0.7221 / 0.3252
150 / 21 / 1,299 / 66.12 / 0.7743 / 0.4185
200 / 20 / 1,357 / 66.36 / 0.7818 / 0.4861
300 / 16 / 80 / 67.74 / 0.7926 / 0.4416
400 / 13 / 88 / 67.05 / 0.8190 / 0.4188
500 / 12 / 90 / 67.37 / 0.8282 / 0.4637
600 / 10 / 136 / 67.22 / 0.8637 / 0.5315
1100 / 9 / 136 / 67.17 / 0.8708 / 0.6966

Part D Plan Reliability (inter-rater agreement)

	Percent Agreement		
Unit of Analysis	**Programmer 1 Num/Den (%)**	**Programmer 2 Num/Den (%)**	**Final Cohen's Kappa**
Part D Plan 1	147/246 (59.8%)	147/246 (59.8%)	1.00
Part D Plan 2	14/32 (43.8%)	14/32 (43.8%)	1.00
Part D Plan 3	52/78 (66.7%)	52/78 (66.7%)	1.00
Part D Plan 4	33/58 (56.9%)	33/58 (56.9%)	1.00
Part D Plan 5	27/50 (54.0%)	27/50 (54.0%)	1.00

Accountable Care Organizations (ACO) Reliability
Minimum denominator size of ACO / # of ACOs / Mean denominator size of ACOs / Mean rate of ACOs / Reliability score (based on the mean rate and the mean denominator size) / Reliability score (based on the mean rate and the minimum denominator size)

60 / 32 / 211 / 66.18 / 0.3744 / 0.2843
100 / 25 / 247 / 66.12 / 0.4763 / 0.3667
150 / 17 / 300 / 67.31 / 0.5536 / 0.4282
200 / 11 / 372 / 66.87 / 0.6060 / 0.3620
250 / 7 / 456 / 66.58 / 0 / 0.4793

370 / 6 / 491 / 66.95 / 0 / 0.6047
380 / 5 / 513 / 66.39 / 0 / 0.6404
410 / 4 / 545 / 66.87 / 0 / 0.7000

Physician Group Reliability
Minimum denominator size of MD group / # of Groups / Mean denominator size of MD group / Mean rate of physician groups at minimum denominator / Reliability score (based on the mean rate and the mean denominator size) / Reliability score (based on the mean rate and the minimum denominator size)

10 / 246 / 24 / 69.09 / 0.4056 / 0.1886
20 / 98 / 39 / 71.27 / 0.5871 / 0.3548
30 / 50 / 54 / 70.95 / 0.6662 / 0.4911
35 / 39 / 60 / 72.21 / 0.7068 / 0.5414
40 / 31 / 66 / 72.76 / 0.7350 / 0.6068
45 / 24 / 72 / 71.66 / 0.7408 / 0.6201
50 / 23 / 74 / 71.08 / 0.7629 / 0.6341
55 / 18 / 80 / 68.90 / 0.7767 / 0.6647
60 / 13 / 88 / 68.83 / 0.8063 / 0.7068
65 / 12 / 90 / 68.21 / 0.8098 / 0.7320

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?*)

State Reliability
We concluded that the reliability test was adequate, since all state-level reliability scores were greater than 0.7, indicating that the measure would produce reliable scores at the state level.

Part D Plan Reliability (signal-to-noise analysis)
Using the method of mean denominator and volume categories, a minimum denominator of 100 and a mean denominator of 1,245 resulted in an overall reliability score of 0.72, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between plans. The aforementioned criteria resulted in 54.6% of plans (12 of 22 plans) with a reliable score. To achieve 100% of plans with a reliable score would require restricting the analysis to plans with a denominator size of 1,100 or greater, as shown by the reliability score (based on the mean rate and the minimum denominator) = 0.6966.

Part D Plan Reliability (inter-rater agreement)
Results obtained by the final two independent programmers were 1.00, which is greater than the Kappa threshold of 0.9. No further refinement of measure specifications was deemed necessary, and the measure specifications are considered reliable.

Accountable Care Organizations (ACO) Reliability
Using the method of mean denominator and volume categories, a minimum denominator of 200 and a mean denominator of 372 had overall reliability of 0.61. This approaches the threshold reliability score of 0.7. To achieve 100% of ACOs with a reliable score would require restricting the analysis to ACOs with a denominator size of 410 or greater, as shown by the reliability score (based on the mean rate and the minimum denominator) = 0.70. In our opinion, reliability scores will improve when measure rates are calculated across all ACOs (n=259), rather than the limited sample (n=32) available for testing.

Physician Group Reliability
Using the method of mean denominator and volume categories, a minimum denominator of 35 and a mean denominator of 60 had overall reliability of 0.71. The aforementioned criteria resulted in 23.1% of physician

groups (9 of 39 physician groups) with a reliable score. To achieve 100% of physician groups with a reliable score would require restricting the analysis to physician groups with a denominator size of 60 or greater, as shown by the reliability score (based on the mean rate and the minimum denominator) = 0.7068.

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

Empirical validity testing is not available for this measure at the time of this maintenance review. Analysis was not possible in the timeframe from NQF publication of this new evaluation criteria (September 2017) and submission of the testing form (January 2018). On March 9, 2018, the measure steward, CMS, met with NQF to discuss submission of this measure. NQF requires empiric validity testing at the time of maintenance; however, they recognize the limitations of the timeframe for submission. NQF, CMS, and the contract team agreed that in leu of providing results of testing, it would be suitable to include a detailed plan for testing empiric validity before the next maintenance submission.

We will test measure performance score validity by examining correlations with meaningful measures of a similar quality construct (convergent validity) using the Spearman's rank correlation coefficient. We will analyze the convergent validity of the measures, evaluating the extent to which the measures *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) and *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) correlate. We hypothesize that health plans and provider groups that perform well at helping individuals with bipolar I disorder remain adherent to mood stabilizers will also perform well at helping individuals with schizophrenia remain adherent to antipsychotic medications. Both measures are indicators of overall quality of care for individuals with serious mental illness and should be correlated.

For health plan level testing, we will evaluate the correlation between *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) and *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) using Medicare-Medicaid Plan (MMP) encounter data. We will begin our initial testing using data already available to us from federal fiscal years 2015 and 2016, covering dates between October 1, 2014, and September 30, 2016. Because of the uncertain quality of the encounter data reported by MMPs, we will conduct an initial series of data checks to examine the quality and volume of encounter data required for the measures and include MMPs for which the quality is sufficient for testing purposes. Our initial data checks will examine quality and volume of data at the plan and state levels to ensure sufficient sample sizes for testing the research questions. We anticipate using data elements related to Medicare and Medicaid enrollment, institutional encounters, non-institutional encounters, and prescription drug coverage and claims.

For provider level testing we will evaluate the correlation between *Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder* (NQF #1880) and *Adherence to Antipsychotic Medications for Individuals with Schizophrenia* (NQF #1879) using Medicare FFS data paired with Medicare Part D claims data. We will pull this Medicare FFS data from the Integrated Data Repository (IDR) to complete testing. No Medicaid data will be used. We anticipate using data elements related to Medicare and Medicaid enrollment, institutional claims, non-institutional claims, and prescription drug coverage and claims.

We will produce scatter plots comparing the two measures at the provider and health plan level. The Spearman's rank correlation coefficient (r_s) assesses the monotonic relationship in plan rankings for each measure pair. The coefficient

ranges from -1 to 1, where $r_s = 1$ indicates perfect alignment of plan rankings, $r_s = -1$ indicates opposite alignment of plan rankings, and $r_s = 0$ represents no alignment in plan rankings. We will fit a smooth curve using locally weighted scatterplot smoothing (LOWESS) method to visualize any trends in the scatterplots. Because the LOWESS method does not rely on a preconceived model for the distribution of the measures (non-parametric), the LOWESS curve can captured detailed information about the measure relationships that the correlation coefficient does not convey.

The timeline for this work is described below:

- October – November 2018: Develop analytic file
- November 2018 – February 2019: Conduct validity testing and review results
- March – April 2019: Summarize results and update measure documentation
- TBD: Submit updated validity testing to NQF as part of maintenance submission

Although empirical validity analysis has not yet been conducted, this measure uses a definition of adherence (0.8 proportion of days covered) that is harmonized with other National Quality Forum (NQF)-endorsed adherence measures and is consistent with the threshold of adherence used in studies cited in the evidence attachment. These studies demonstrated improved outcomes in bipolar I associated with adherence to medication. Although many of these studies have used the medication possession ratio (MPR) rather than the proportion of days covered (PDC), CMS and the Pharmacy Quality Alliance have evaluated and extensively tested the PDC and the MPR and specifically found that: 1) the PDC and MPR will provide nearly identical results when examining adherence to a single drug; 2) the PDC will provide a more conservative estimate of adherence when examining adherence to a class of drugs that are prone to frequent switching and concomitant therapy with multiple drugs within the class (as with antipsychotic drugs). Therefore, based on NQF's recommendation that a standard methodology for calculating medication adherence be established across all endorsed adherence measures, CMS and PQA agreed to harmonize the methodology for calculating medication adherence using the PDC, which was approved by the NQF Consensus Standards Approval Committee (CSAC).

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

A Technical Expert Panel (TEP), comprised of internal medicine physicians and pharmacists, evaluated the face validity of the measure and measure scores.

The 14 member TEP included the following individuals listed below. Ultimately, 12 of the 14 TEP members evaluated the face validity of the measure and the measure scores.

1. Jill S. Borchert, PharmD, BCPS, Professor, Pharmacy Practice & PGY1 Residency Program Director, Midwestern University, Chicago College of Pharmacy
2. Anne Burns, RPh, Vice President, Professional Affairs, American Pharmacists Association
3. Jannet Carmichael, PharmD, FCCP, FAPhA, BCPS, VISN 21 Pharmacy Executive, VA Sierra Pacific Network
4. Marshall H. Chin, MD, MPH, Professor of Medicine, University of Chicago
5. Jay A. Gold, MD, JD, MPH, Senior Vice President and Medicare Chief Medical Officer, MetaStar, Inc.
6. David Nau, PhD, RPh, CPHQ, Senior Director of Research & Performance Measurement, PQA, Inc.
7. N. Lee Rucker, MSPH, Senior Strategic Policy Advisor, AARP - Public Policy Institute
8. Marissa Schlaifer, MS, RPh, Director of Pharmacy Affairs Academy of Managed Care Pharmacy
9. Brad Tice, PharmD, Chief Clinical Officer, PharmMD Solutions, LLC

10. Jennifer K. Thomas, PharmD, Manager, Pharmacy Services, Delmarva Foundation for Medical Care/Delmarva Foundation of the District of Columbia
11. Darren Triller, PharmD, Director, Pharmacy Services, IPRO
12. Neil Wenger, MD, Professor of Medicine, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research
13. Edward Eisenberg, Vice President and Chief Medical Officer, Medicare, Medco Health Solutions; Franklin Lakes, NJ
14. Douglas Bell, Associate Professor in Residence, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research; Los Angeles, CA

The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using Survey Monkey). Face validity of the measure score as an indicator of quality was systematically assessed as follows: After the measure was fully specified and tested, the expert panel members were asked to rate, based on a 5-point Likert scale, their level of agreement with the following statement: "The measure appears to measure what is intended."

The 5-point Likert scale was defined as follows: 1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree

ICD-10-CM Conversion Methodology
The conversion of the measure to include ICD-10-CM codes is provided as requested by NQF. The crosswalk is provided as an excel file in Section S2.b Data Dictionary or Code Table.

Name and Credentials of Experts Who Assisted in the Process

- Soeren Mattke, MD, DSc, Senior Scientist, RAND Corporation
- Tim Laios, MBA, MPH, Executive Director, Informatics, Health Services Advisory Group (HSAG)
- Ryan Fair, BS, Director, Informatics, HSAG
- Kerri Carlile, MS, Informatics Analyst, HSAG
- Sara Lomeli, BA, Informatics Project Coordinator, HSAG

Evaluation of ICD-9-CM Changes
The changes (i.e., deletions and/or additions) made to the ICD-9-CM codes for the measures requiring conversion were reviewed. Additionally, the ICD-9-CM codes were reviewed for any coding updates, using the October 2012 Conversion Table of New ICD-9-CM Codes, published by the National Center for Health Statistics (NCHS) and the Centers for Medicare & Medicaid Services (CMS).

ICD-9-CM Code Identification
For each measure requiring conversion, original tables were used to identify all ICD-9-CM codes included in the measure. Those ICD-9-CM codes and matching descriptions were then extracted from the Ingenix 2011 ICD-9-CM Data File. Only valid ICD-9-CM codes were retained and used in the ICD-9-CM to ICD-10-CM conversion process.

Ingenix Extraction
When extracting the ICD-9-CM codes from the Ingenix Data File, all codes were extracted with two-decimal specificity. For example, for ICD-9-CM code 274.1, all ICD-9-CM codes that had 2741 for the first four digits were extracted (e.g., 274.10, 274.11, and 274.19). For every three-digit ICD-9-CM code used in the measure, all ICD-9-CM codes that began with those first three digits were extracted. For the ICD-9-CM codes listed in ranges, codes with up to two-decimal specificity were extracted within that range.

Conversion Process

The ICD-9-CM and ICD-10-CM General Equivalence Map (GEM) text files and the ICD-10-CM Descriptions text file were imported into SAS and combined into one data file to capture all ICD-9-CM codes, their corresponding ICD-10-CM codes, and the ICD-10-CM code descriptions. The ICD-9-CM codes that were retained from the Ingenix 2011 ICD-9-CM Data File described above were then extracted from the combined GEM data file.

The results for each measure were then exported into Excel and validated to ensure that the translation was appropriate (i.e., the crosswalk was correct and applied appropriately and all appropriate ICD-9-CM codes were captured). Since one ICD-9-CM code can have several corresponding ICD-10-CM codes, each ICD-9-CM code can have multiple entries in the final Excel document (i.e., one row for each corresponding ICD-10-CM code).

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

Since face validity was used, the systematic assessment was conducted, and the results are described below:

Systematic Assessment of Face Validity
The results of the Technical Expert Panel rating of face validity as represented by this statement, "The measure appears to measure what is intended," on a scale of 1 to 5 are presented here:
N=12 panel members, Mean Rating=4.08

Response / % of TEP / Number of TEP
5=Strongly Agree / 16.7% / 2
4=Agree / 75.0% / 9
3=Neutral / 8.3% / 1
2=Disagree / 0.0% / 0
1=Strongly Disagree / 0.0% / 0

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?*)

In summary, 11 of the 12 TEP members either responded “agree” or “strongly agree” that the measure, as specified, exhibited face validity. A single TEP member voted “neutral” and requested a change to the measure description, which was incorporated and therefore also supported the face validity of the measure. These responses indicate strong support of the face validity of the measure by the Technical Expert Panel.

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** 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)*

To identify statistically significant differences in performance, we conducted a comparison of means and percentiles at the state, Part D plan, ACO, and physician group level. Confidence intervals (95%) were calculated around point estimates for each state, Part D plan, ACO, and physician group, and then compared to the overall mean of states, Part D plans, ACOs, and physician groups respectively. If the confidence intervals did not overlap with the overall mean, the difference was considered statistically significant.

Furthermore, for health plans, the observed sample sizes of members of each comparison unit were tested empirically to determine whether there was sufficient power to detect statistically significant differences between members (e.g., between plans). To do this, all members were divided into quintiles according to their measure score. Within each quintile, the member with a denominator closest in size to the median denominator of the quintile and the member with the measure score closest to the median measure score of that quintile were identified. Comparison of members based on their median denominator size was made, because a relationship between denominator size and quality cannot be excluded a priori. In addition, a "standardized" member of each quintile was simulated by using the median denominator size across all quintiles. Binomial (exact) 95% confidence intervals for each of the 10 selected plans (i.e., two plans per quintile) were calculated around the point estimates. Overlapping confidence intervals indicate insufficient statistical power to detect statistically significant differences.

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 analyzed the measure performance by state, Part D plan, ACO, and physician group, and the results, along with a discussion of the meaningful differences at each level, are described below:

Meaningful Differences at the State Level

Below we present the measure rate by state, mean, median, and standard deviation.

AZ – 60.8%
DE – 63.8%
FL – 64.6%
IA - 77.4%* (statistically significantly higher than the mean)
IN – 70.1%* (statistically significantly higher than the mean)
MS – 61.0%
RI – 69.6%
WA – 71.8%* (statistically significantly higher than the mean)
Mean of state scores– 67.4%
Median of state scores - 67.1%
Standard Deviation of state scores – 5.8%

Meaningful Differences at the ACO Level
Below we present the mean, standard deviation, and percentiles at the ACO level.

Number of ACOs with minimum denominator of at least 30 in the denominator = 32.
Mean: 66.2%
SD: 5.8%
10th Percentile: 58.1%
25th Percentile: 63.2%
50th Percentile: 67.3%
75th Percentile: 69.3%
90th Percentile: 74.2%
Of the ACO scores, 4/32 (12.5%) of providers were statistically significantly lower than the mean; 3/32 (9.4%) of providers were statistically significantly higher than the mean.

Meaningful Differences at the Physician Group Level
Below we present the mean, standard deviation, and percentiles at the physician group level.

Number of physician groups with at least 30 in the denominator = 50.
Mean: 71.0%
SD: 10.5%
10th Percentile: 54.6%
25th Percentile: 64.3%
50th Percentile: 72.7%
75th Percentile: 78.1%
90th Percentile: 83.7%
Of the physician group scores, 3 out of 50 (6.0%) of providers were statistically significantly lower than the mean and 16 out of 50 (32.0%) of providers were statistically significantly higher than the mean, indicating a wide range of scores.

Meaningful Differences at the Health Plan Level
Below we present the mean, standard deviation, and percentiles at the Part D plan level.

Number of Plans with at least 30 in the denominator= 27
Mean: 66.5%
SD: 5.1%
10th Percentile: 61.0%

25th Percentile: 63.4%
50th Percentile: 66.4%
75th Percentile: 69.6%
90th Percentile: 72.6%
Of the plans with at least 30 in the denominator, 3/27 (11.1%) of providers were statistically significantly lower than the mean, 11.1% of providers were statistically significantly higher than the mean.

Across Part D Plan with ≥ 30 Beneficiaries	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Number of plans	4	5	5	5	4
Denominator range across plans (minimum-maximum)	47-129	212-1,288	81-1,747	150-1,932	42-1,270
Median denominator size per plan	105	623	1,059	271	139
Measure score (95% CI) of the plan with a denominator size closest to the median denominator size	50.9% (42.2-60.3)	59.1% (55.3-63.0)	62.0% (59.2-65.0)	63.8% (58.3-69.6)	66.7% (58.3-75.3)
Measure score range across plans	48.9%-57.0%	58.5%-60.1%	60.3%-62.1%	62.7%-63.8%	63.9%-69.1%
Median measure score	51.4%	59.8%	61.7%	63.4%	66.3%
95% CI of the plan with a score closest to the median score	50.9% (42.2-60.3)	59.8% (56.3-63.5)	61.7% (51.7-72.3)	63.4% (56.3-70.8)	66.7% (58.3-75.3)
95% CI based on the overall median denominator size N=212	51.4% (44.9-58.3)	59.8% (53.4-66.5)	61.7% (55.4-68.3)	63.4% (57.1-69.9)	66.3% (60.2-72.7)
CI = Confidence Interval					

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?*)

Meaningful Differences at the State Level
Three of the eight states (37.5%) had scores statistically significantly lower than the mean and another three states had scores significantly higher than the mean. Measure rates ranged from 60.8% in AZ to 77.4% in IA, indicating suboptimal performance across all 8 states.

Meaningful Differences at the ACO Level
Of the ACO scores, 4/32 (12.5%) of providers were statistically significantly lower than the mean; 3/32 (9.4%) of providers were statistically significantly higher than the mean.

For those ACOs with at least 30 eligible individuals, high- (90th percentile) and low- (10th percentile) performing ACO were 16.1% apart, indicating suboptimal performance across all ACOs and variation between high- and low-performing ACOs.

Meaningful Differences at the Physician Group Level

For those physician groups with at least 30 eligible individuals, high- (90th percentile) and low- (10th percentile) performing physician groups were 29.1% apart. The results indicate ample room for improvement and meaningful differences in quality of care between the highest and lowest performing physician groups.

Meaningful Differences at the Health Plan Level

Of the plans with at least 30 in the denominator, 3/27 (11.1%) of providers were statistically significantly lower than the mean, 11.1% of providers were statistically significantly higher than the mean.

For those plans with at least 30 eligible individuals, high- (90th percentile) and low- (10th percentile) performing plans were 11.6% apart, indicating suboptimal performance across all plans and variation between high- and low-performing plans.

A total of 23 plans with at least 30 beneficiaries could be distributed across the measure score quintiles. Plans showed pronounced variation in sample size with a general pattern in the first 4 quintiles of increasing size with respect to measure scores. Comparison of standardized plans (with confidence intervals calculated based on the overall median denominator size of the entire sample) showed sufficient discriminatory ability between members of the highest and lowest quintiles. Of note, the sample sizes for plans varied dramatically within each quintile and will result in distinctly different power if two plans are compared.

Assuming a median measure rate of 61.7% and a median denominator of 212 beneficiaries, the smallest difference in measure rates that can be detected at the plan level with a power of 80% and α=0.05 is 9.6%.

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 days' supply data and bias from cash prescriptions were possible threats to validity. An empirical assessment of these possible threats was conducted as follows:

Threat of Bias from Missing Data
We have identified two potential scenarios where measure results could be biased by missing data:

1. Missing days' supply within the prescription drug event data, which is a required data element to calculate medication adherence;
2. Missing drug claims due to individuals using alternative payment sources for prescription drugs, e.g., $4 commercial discount prescription programs and other alternative drug benefits, such as the Veterans Administration (VA)

For missing days' supply, we analyzed the number (%) of beneficiaries in the measure denominator with one or more claims that had missing days' supply.

For bias from cash prescriptions or alternative sources, we conducted a limited sensitivity analysis using a two-state sample (FL and RI) to estimate the potential impact of a commercial cash discount program on measure rates. Specifically, we created a National Drug Code (NDC) list from the formulary of a leading cash discount program to identify those individuals with at least one claim for a mood stabilizer on the formulary and no claims for any other Part D drugs on the formulary as a proxy to potentially indicate the individual was filling medications through the cash discount program. We then simulated the effect on measure rates, if each of these individuals' mood stabilizer medication use extended from the last consecutive claim to the end of the measurement period, assuming that individuals had switched to the cash program. We simulated two scenarios: including complete coverage of all remaining days until the end of the measurement period was 100% or extrapolating the average proportion of days covered from the first prescription in the measurement period to the last prescription in the measurement period.

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*)

Missing Data
Days' Supply: Only 3 individuals (0.01%) in the overall measure denominator had one or more claims with missing days' supply.

Cash Prescriptions
The percentage of individuals in the denominator with mood stabilizer Part D claims on the formulary and no claims for any other drugs on the commercial discount formulary was 1.9% (219/11,575).

SCENARIO 1. If individuals with possible cash prescriptions (i.e., those described above) are assumed to have mood stabilizer medication for all days from the last day covered to the end of the measurement period (i.e., 100% adherence), the PDC would be (63.4%) (7,337/11,575).

SCENARIO 2. If individuals with possible cash prescriptions (i.e., those described above) are assumed to have antipsychotic medication for all days from the last day covered at the same proportion as the PDC calculated over the period from first to last claim in the measurement period (i.e., same adherence as the rest of the period), the PDC would be 7,343/11,575 (63.4%) (FL and RI only)

The actual measure rate was 7,337/11,575 (63.4%) (FL and RI only).

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
Only 3 individuals (0.01%) in the overall measure denominator had one or more claims with missing days' supply. This small number of cases indicates that missing data do not pose a threat to the validity of the measure.

Cash Prescriptions
The actual measure rate was 7,337/11,575 (63.4%) (FL and RI only). Therefore, the findings suggest that very little impact on measure rates would be expected from utilization of the cash discount program. Of note, this analysis is exploratory in nature and assumes that individuals were not switched to a drug on the commercial discount formulary, and if they were utilizing the discount program, they were obtaining all of their medications at a cash discount program. Additional limitations include prescriptions filled with other benefits (e.g., VA), and the extent to which this measure might underestimate mood stabilizer use due to those factors is unknown.

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.
Testing demonstrated that the data required were available and accessible. Issues affecting feasibility regarding missing data were not identified. The cost of data collection is negligible, since the administrative data (collected by CMS primarily for billing purposes) are used as the data source for this measure. Other feasibility/implementation issues were not identified.

DATA COLLECTION
Testing was conducted with the CMS administrative claims data. No additional data collection was conducted.

AVAILABLILITY OF DATA
Testing was conducted with the CMS administrative claims data. The data were readily available and accessible.

MISSING DATA
No threats to the validity of this measure were identified using a limited analysis designed to address missing data (Reference Validity Testing Section 2b2.2).

TIMING AND FREQUENCY OF DATA COLLECTION
Testing was conducted with the CMS administrative claims data. Data sources needed to implement the measure are collected by CMS in a timely manner.

SAMPLING
Not Applicable

PATIENT CONFIDENTIALITY
Not Applicable

TIME AND COST OF DATA COLLECTION
The administrative data (collected by CMS primarily for billing purposes) are used as the data source for this measure. Therefore, the cost of data collection is negligible.

OTHER FEASIBLITY/IMPLEMENTATION ISSUES
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)*.
Proprietary coding is contained in the attached list of codes. Users of the proprietary code sets should obtain all necessary licenses from the owners of these code sets.

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 Quality Improvement (Internal to the specific organization)	Payment Program New York State Delivery System Reform Incentive Payment (DSRIP) Program https://www.health.ny.gov/health_care/medicaid/redesign/dsrip/vbp_library/quality_measures/docs/2018_harp_qms.pdf

Not in use	Quality Improvement (external benchmarking to organizations) Substance Abuse and Mental Health Services Administration (SAMHSA) section 223 demonstration https://www.samhsa.gov/sites/default/files/programs_campaigns/ccbhc-criteria.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

New York State Delivery System Reform Incentive Payment (DSRIP) Program: The measure is publicly reported (though not required) in New York State's Delivery System Reform Incentive Payment (DSRIP) Program, and is included in the Value Based Payment (VBP) Quality Measure Set for the Health and Recovery Plan (HARP) subpopulation. As of 2016, 45,000 individuals were enrolled in HARP. HARP is a specialized managed care program for adult individuals with Severe Mental Illness (SMI) or Substance Use Disorder (SUD) that began its rollout in New York State on October 1, 2015. For HARP, the VBP pilot was implemented in two health plans at two different providers. This measure was selected as clinically relevant, reliable, valid, and feasible; however, it is currently not required to report. Pay for reporting measures are intended to be used by the Managed Care Organizations (MCOs) to incentivize VBP Contractors for reporting data to monitor quality of care delivered to members under a VBP contract. Incentives for reporting should be based on timeliness, accuracy, and completeness of data.

Substance Abuse and Mental Health Services Administration (SAMHSA) Section 223 Demonstration Program: This program is authorized under Section 223 of the Protecting Access to Medicare Act (PAMA). Program activities aim to integrate behavioral health with physical health care, increase consistent use of evidence-based practices, and improve access to high-quality care. Participating states in the demonstration program certify community behavioral health clinics that meet federally developed criteria emphasizing accessible and high-quality care. The certified community behavioral health clinics (CCBHCs) are compensated for services through a prospective payment system (PPS).

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?*)

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.*)

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.
New York State DSRIP Program: This measure began being piloted for the HARP subpopulation in 2017 with results being reported (though not required) in 2018. The New York State Department of Health website provides a library of resources for providers and health plans including the technical specifications manual, webinars, and information about the advisory groups involved. The state also holds workshops to explain the VBP process and expectations.

SAMHSA Section 223 Demonstration Program: In 2015, the Department of Health and Human Services (HHS) awarded CCBHC planning grants (Phase I) to 24 states, and eight of those states were selected to participate in the demonstration program (Phase II) to improve access to high-quality behavioral health programs. The CCBHC demonstration program and PPS are designed to work within the scope of state Medicaid Plans and to apply specifically to individuals who are Medicaid enrollees. The eligible population in these states includes all behavioral health clinic (BHC) consumers served by a BHC provider.

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.**

New York State DSRIP Program: This measure is not required to be reported. Information on the process are provided in New York State's, 2018 Value Based Payment Reporting Requirements Technical Specifications Manual. Medicaid Managed Care Organizations with Level 1 or higher value–based contracting arrangements or MCOs with a VBP Pilot contract are required to report. Plans will electronically submit patient-level detail files and patient attribution files via secure file transfer on August 1, 2018. New York State provides VBP contractors and MCOs with a dynamic data and analytics tool that provides cost and outcome information of the different VBP arrangements, by MCO, by geography and by provider(s), including potentially shared savings.

SAMHSA Section 223 Demonstration Program: Certified community behavioral health clinics and their states are required to collect 21 of 32 quality measures for the demonstration program. This measure is not required to be reported. For each demonstration year (the measurement year), quality measures and metrics are submitted within nine months for CCBHCs, and within 12 months for states. CCBHC-lead data and measures are reported to their designated state agency, and state-lead data and measures are reported to SMAHSA by email. SAMHSA will share the data with CMS for the purposes of Quality Bonus Payments and with the Office of the Assistant Secretary for Planning and Evaluation (ASPE) for the purposes of evaluation. Data is reported by using the data reporting templates, and relaying on the major specifications and instructions for those templates found in the Technical Specifications and Resource Manual. SAMHSA's technical assistance (e.g. webinars, guidance documents) is designed to help states and clinics collect, analyze and report the data for each measure. Clarifications related to quality measures and data reporting are provided on the SAMHSA website, and additional questions are submitted by email.

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.
New York State DSRIP Program: The program is in its first pilot year and performance has not yet been reported. The state receives feedback on quality measure feasibility, reporting, and calculation from a VBP Measure Support Task Force, including professionals from various Managed Care Organizations (MCOs), VBP Pilot Contractors, State Agencies, along with other professionals with experience in quality measurement and health information technology. They also receive input from a Clinical Advisory Group that evaluates feedback from VBP Contractors, MCOs, and stakeholders, any significant changes in evidence base of underlying measures and/or conceptual gaps in the measurement program. Feedback from these groups is not publicly available at this time.

SAMHSA Section 223 Demonstration Program: For the purposes of continuous quality improvement, behavioral health clinics (BHCs) submit data and measure results to the state. Ongoing refinement of the system at both the state and BHC level is achieved through state feedback to the BHC regarding the data and measure results, and BHC internal feedback and adjustment regarding both data and results. Feedback from these groups is not publicly available at this time.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
New York State DSRIP Program: No feedback specific to this measure is currently available.

SAMHSA Section 223 Demonstration Program: No feedback specific to this measure is currently available.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure recently went through a re-evaluation process. During that process, feedback on the measure was obtained from measure advisory panels including NCQA's Pharmacy Panel and NCQA's Behavioral Health Measure Advisory Panel. These panels recommended adding medications which are FDA approved for the treatment of bipolar I disorder and removing medications which are not FDA approved for the treatment of bipolar I disorder.

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.**
Based on the feedback obtained from NCQA's Pharmacy Panel and NCQA's Behavioral Health Measure Advisory Panel (described 4a2.2.3) the following measure changes were implemented:

1. Add the following FDA approved medications to the measure as recommended by the pharmacy panel and BHMAP:
- Cariprazine
- Quetiapine fumarate (Seroquel)

2. Remove the following off-label medications from the measure as recommended by the pharmacy panel and internal review of FDA labels (these medications were included in the original measure specification):
- Fluphenazine

- Haloperidol
- Molindone
- Perphenazine
- Pimozide
- Prochlorperazine
- Thioridazine
- Thiothixene
- Trifluoperazine
- Clozapine
- Iloperidone
- Paliperidone
- Fluphenazine decanoate
- Haloperidol decanoate
- Olanzapine pamoate
- Paliperidone palmitate

3. Add the following code to the value set for identifying bipolar I disorder in the measure: F30.8 (other manic episodes).

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.
New York State DSRIP Program: Performance data is not publicly available for this measure.

SAMHSA Section 223 Demonstration Program: Performance data is not publicly available for this measure.

We envision this measure will help providers to identify patients with bipolar I disorder who are not adherent (at a critical threshold of 0.8 or greater) with long-term treatment with mood stabilizer medications and target interventions to improve medication adherence.

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 were no identified unintended findings for this measure during testing and none have been brought to our attention since implementation.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
No unexpected benefits.

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)
0003 : Bipolar Disorder: Assessment for diabetes
0109 : Bipolar Disorder and Major Depression: Assessment for Manic or hypomanic behaviors
0110 : Bipolar Disorder and Major Depression: Appraisal for alcohol or chemical substance use
0111 : Bipolar Disorder: Appraisal for risk of suicide
0112 : Bipolar Disorder: Level-of-function evaluation
0541 : Proportion of Days Covered (PDC): 3 Rates by Therapeutic Category
0542 : Adherence to Chronic Medications
0543 : Adherence to Statin Therapy for Individuals with Cardiovascular Disease
0545 : Adherence to Statins for Individuals with Diabetes Mellitus
0580 : Bipolar antimanic agent
1879 : Adherence to Antipsychotic Medications for Individuals with Schizophrenia
1927 : Cardiovascular Health Screening for People With Schizophrenia or Bipolar Disorder Who Are Prescribed Antipsychotic Medications
1932 : Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)

5.1b. If related or competing measures are not NQF endorsed please indicate measure title and steward.
Adherence to Antipsychotic Medications for Individuals with Schizophrenia. NCQA is measure 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 specifications are harmonized with the related measure, Adherence to Antipsychotic Medications for Individuals with Schizophrenia (NQF #1879) and the NCQA version of the same measure (Adherence to Antipsychotic Medications for Individuals with Schizophrenia), where possible. The methodology used to calculate adherence in these measures is proportion of days covered (PDC) which is calculated the same in all three measures. The methodology used to identify the denominator population is also calculated the same in all three measures, with the exception of the clinical conditions which is the target of the measure. The data collection burden is identical for the measures. The only differences between Adherence to Mood Stabilizers for Individuals with Bipolar I Disorder (NQF #1880), Adherence to Antipsychotic Medications for Individuals with Schizophrenia (NQF #1879), and the related NCQA measure are: (1) the clinical codes used to identify the different populations in each measure (NQF #1880 – individuals with bipolar I disorder; NQF #1879 and NCQA measure– individuals with schizophrenia); (2) the medications includes in each measure (NQF #1880- mood stabilizers; NQF #1879 and the NCQA measure– antipsychotics); and, (3) an exclusion for dementia which is included in NQF #1879 and the NCQA measure but not in NQF #1880. The rationale for these difference is due to the different clinical focus of each measure. There is no impact on interpretability since the measures clearly identify the disparate clinical focus. During development the measure developers worked to harmonize this measure with other measures which were NQF-endorsed at the time of development. The section below is from the original submission of the measure for initial endorsement and refers to measures which are no longer NQF-endorsed. We are including this language to demonstrate the efforts of the measure developers to harmonize this measure with other measures. MEASURES WITH WHICH THE MEASURE IS HARMONIZED. The measure has been harmonized where feasible with NQF #0542, #0543, #0545, #0541, #1879, #1927, and #1932 MEASURES WITH WHICH THE MEASURE IS NOT HARMONIZED. The measure specifications of the measure are not harmonized with the following NQF-endorsed measures that have the same measure focus (use of mood stabilizers among patients with Bipolar Disorder): NQF #0580 Bipolar antimanic agent. DIFFERENCES BETWEEN MEASURE 1880 AND MEASURE 0580. One NQF-endorsed measure (NQF #0580) focuses on a similar concept, but differs from this measure in two important

ways. First, the NQF-endorsed measure includes individuals with newly diagnosed bipolar disorder and major depressive disorder. However, this measure includes all individuals with bipolar I disorder, not just those who are newly diagnosed, and does not include individuals with major depressive disorder. Second, the NQF-endorsed measure identifies the percentage of eligible individuals who have received at least 1 prescription for a mood-stabilizing agent during the measurement year, while this measure measures the percentage of eligible individuals with a proportion of days covered (PDC) for mood stabilizer medications greater than 0.8 during the measurement year. RATIONALE. This measure is an improved measure that adds value because it measures adherence to mood stabilizer treatment for individuals with bipolar I disorder. In contrast, the NQF measure (NQF# 0580) is linked to a one-time prescription for mood stabilizer treatment. IMPACT ON INTERPRETABILITY AND DATA COLLECTION BURDEN. Differences have not been identified concerning the data collection burden between Measure 1880 and Measure 0580. However, interpretability for Measure 1880 (as compared to NQF #0580) is improved because Measure 1880 focuses on adherence rather than a single prescription, and Measure 1880 is harmonized with the majority of adherence measures for other chronic diseases in the NQF portfolio and those that are being publicly reported by CMS.

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 does not address both the same measure focus and population as another NQF-endorsed measure.

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
Co.2 **Point of Contact:** Elizabeth, Ricksecker, Elizabeth.Ricksecker@cms.hhs.gov, 410-786-6723-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, swift@ncqa.org, 202-955-5174-

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.
Behavioral Health Measure Advisory Panel (BHMAP) – advised on the re-evaluation:

1. Katherine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute
2. Christopher Dennis, MD, MBA, FAPA, Chief Behavioral Health Officer, Landmark Health
3. Ben Druss, MD, MPH, Professor, Emory University
4. Frank A. Ghinassi, PhD, ABPP, President and CEO, Rutgers University Behavioral Health Care
5. Connie Horgan, ScD, Professor and Director, Institute for Behavioral Health, Brandeis University
6. Laura Jacobus-Kantor, PhD, Chief, Quality, Evaluation and Performance, SAMHSA HHS
7. Jeffrey Meyerhoff, MD, National Medical Director for Medicare and Retirement, Optum Behavioral Solutions
8. Harold Pincus, MD, Professor and Vice Chair--Department of Psychiatry, College of Physicians and Surgeons, Co-Director, Irving Institute for Clinical and Translational Research, Columbia University, Director of Quality and Outcomes Research, New York –Presbyterian Hospital
9. Michael Schoenbaum, PhD, Senior Advisor for Mental Health Services, Epidemiology and Economics, National Institute of Mental Health

10. John Straus, MD, Medical Director Special Projects, Massachusetts Behavioral Health Partnership A Beacon Health Options Company
11. William Wood, MD, PhD, Manager, Medical Director Behavioral Health, Anthem, Inc.

HEDIS Expert Pharmacy Panel – advised on the re-evaluation:
1. Linda DeLaet, PharmD, Kaiser Permanente
2. Gerry Hobson, RPh, Cerner Multum
3. Chronis H. Manolis, RPh, UPMC Health Plan
4. Cathrine Misquitta, PharmD, MBA, BCPS, CGP, FCSHP, Health Net Pharmaceutical Services
5. Kevin Mark, MD, Wisconsin First, Inc.

FMQAI (now HSAG) TEP - advised on the original measure development and testing:

1. Jill S. Borchert, Pharm.D., BCPS, Professor, Pharmacy Practice and PGY1 Residency Program Director, Midwestern University, Chicago College of Pharmacy
2. Anne Burns, RPh, Vice President, Professional Affairs, American Pharmacists Association
3. Jannet Carmichael, Pharm.D., FCCP, FAPhA, BCPS, VISN 21 Pharmacy Executive, VA Sierra Pacific Network
4. Marshall H. Chin, MD, MPH, Professor of Medicine, University of Chicago
5. Jay A. Gold, MD, JD, MPH, Senior Vice President and Medicare Chief Medical Officer, MetaStar, Inc.
6. David Nau, Ph.D., R.Ph., CPHQ, Senior Director of Research and Performance Measurement, PQA, Inc.
7. N. Lee Rucker, M.S.P.H., Senior Strategic Policy Advisor, AARP - Public Policy Institute
8. Marissa Schlaifer, MS, RPh, Director of Pharmacy Affairs Academy of Managed Care Pharmacy
9. Brad Tice, Pharm.D., Chief Clinical Officer, PharmMD Solutions, LLC
10. Jennifer K. Thomas, Pharm.D., Manager, Pharmacy Services, Delmarva Foundation for Medical Care/Delmarva Foundation of the District of Columbia
11. Darren Triller, Pharm.D., Director, Pharmacy Services, IPRO
12. Neil Wenger, MD, Professor of Medicine, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research
13. Edward Eisenberg, Vice President and Chief Medical Officer, Medicare, Medco Health Solutions; Franklin Lakes, NJ
14. Douglas Bell, Associate Professor in Residence, UCLA Department of Medicine, Division of General Internal Medicine and Health Services Research; Los Angeles, CA

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:**

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 04, 2019

Ad.6 **Copyright statement:** Not Applicable

Ad.7 **Disclaimers:** Not Applicable

Ad.8 **Additional Information/Comments:** Not Applicable

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 1932
Measure Title: Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18 – 64 years of age with schizophrenia or bipolar disorder, who were dispensed an antipsychotic medication and had a diabetes screening test during the measurement year.
Developer Rationale: As patients with schizophrenia or bipolar disorder are at an increased risk for diabetes, and antipsychotic medications are an expected treatment that increases the risk of metabolic diseases, screening for diabetes will allow for proper diagnosis and treatment, if warranted.

Numerator Statement: Among patients 18-64 years old with schizophrenia or bipolar disorder, those who were dispensed an antipsychotic medication and had a diabetes screening testing during the measurement year.
Denominator Statement: Patients ages 18 to 64 years of age as of the end of the measurement year (e.g., December 31) with a schizophrenia or bipolar disorder diagnosis and who were prescribed an antipsychotic medication.
Denominator Exclusions: Exclude members who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.
Exclude patients with diabetes during the measurement year or the year prior to the measurement year.

Exclude patients who had no antipsychotic medications dispensed during the measurement year.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System, Population : Regional and State

Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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 which shows that patients with schizophrenia or bipolar disorder are at an increased risk for diabetes. Therefore, screening for diabetes provides an opportunity for early diagnosis and treatment and may reduce poor health outcomes.
- The developer provides a systematic review of the evidence including:
    - American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition. **Recommendations within these guidelines range from Grade I (substantial clinical confidence) to Grade II (moderate clinical confidence).**
    - American Diabetes Association (2018). Standards of medical care in diabetes--2018. **Grade B (supportive evidence from well-conducted cohort studies)**
    - Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. (2016). **This systematic review was conducted in accordance with the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard.**
- In addition, the developer cites the APA 2009 Guideline Watch which cites additional RCTs and studies that have been completed since the 2004 APA Practice Guidelines for the Treatment of Patients with Schizophrenia that furthers the known link between metabolic side effects and antipsychotics used to treat schizophrenia.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided additional systematic reviews of evidence listed above.

Exception to evidence:

N/A

Questions for the Committee:

- *The evidence provided by the developer is updated and directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *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: high; Consistency: high (Box 5) > High (Box 5a) > High

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 summarized the performance data at the health plan level using HEDIS health plan performance rates from 2015-2017. The data is stratified by year and insurance type.

Measurement Year	# of Plans	Median Denom. Size per plan	Mean	St Dev	10th	25th	50th	75th	90th	Interquartile range
2015	143	437	79.8%	0.1	72.7%	75.7%	80.1%	83.8%	87.0%	8.1
2016	185	804	80.4%	0.1	72.3%	77.4%	80.7%	84.0%	87.2%	6.6
2017	202	1,018	80.7%	0.1	74.0%	77.5%	81.0%	84.2%	87.4%	6.7

- In the previous review of this measure (2012) the developer provided field tests results to show a performance gap. Among 22 states, the measure had a minimum value of 2.3%, mean=12.1%, 25th percentile=8.4%, median=10.3%, 75th percentile=16.7% and a maximum value of 28.2%.

Disparities

- The developer does not provide disparities data since HEDIS data is stratified by type of insurance. While not specified in this measure, this measure can also be stratified by demographic variables in order to assess the health care disparities.
- The developer provides a summary of research studies demonstrating that individuals with serious mental illness have an increased risk for diabetes as well as disparities in their care.

Questions for the Committee:

o *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:

**The developer provides a logic model which shows that individuals with schizophrenia and bipolar disorder are at increased risk for diabetes and provides an updated systematic review of the evidence including practice guidelines and meta-analyses.

**The causal pathway is clear and makes sense. The ultimate tie to patient oriented outcomes is face valid, but unproven.

**Evidence is abundant about the increased risk of diabetes for patients with schizophrenia who are also being medicated. A number of new systematic reviews and RCTs are presented as well as inclusion in national standards of the AmerDiabAssoc and a new large-scale meta analysis. Evidence is rated high.

**The developers cite the guidelines from the APA and ADA to support universal glucose testing but those guidelines do not recommend universal testing. They only recommend that clinicians consider glucose testing, among other types of tests that should be considered. Also, they point out that certain populations at higher risk should be tested, such as those who are obese.

1b. Performance Gap

Comments:

**The developer demonstrates a continued performance gap--with the 90th percentile performing at 87.4% and the 10th percentile performing at 74%.

**There is little if any evidence of improvement, although there does appear to be some evidence of gap. No evidence for disparities is noted and this should be a very high priority, if not a requirement. The literature cited does not adequately answer the concerns in my opinion.

**The performance gao on HEDIS between 2015 and 2017 is highest between the 10th (74%) and the 90th percentile (87.4%). For the Medicaid population there has been a 3% improvement in 6 years suggesting the need for performance incentives for this population if improvement is expected.

**The mean testing rate is 81%. Given that testing is not recommended universally, this seems like a high rate and that there is not a performance gap.

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.

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)?*
- *The 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 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**

Committee pre-evaluation comments

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

2a1. Reliability – Specifications
Comments:
**Data elements are clearly defined.
**Clearly defined and unlikely to be prone to unreliability.
**No concerns.

2a2. Reliability – Testing
Comments:

**No--specifications are clear and testing was conducted using the beta binomial method with a result of 0.959.
**High reliability.
**No concerns.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:
**Two methods for conducting validity testing were employed: construct validity with measure on diabetes monitoring for individuals with diabetes and schizophrenia or bipolar disorder with a pearson correlation of 0.25 and face validity multi stake-holder advisory panels and public comment. There was not missing data.
**Moderate given lack of score level testing.
**No concerns.
**Validity was tested by examining the correlation between plans that scored high on this measure, and scores that score high on diabetes monitoring for people with schizophrenia. This could simply reflect the fact that those plans have providers that do a lot of testing in general, not higher quality of care. Ideally, one would want to see if the measure was associated with better outcomes (e.g., lower hyperglycemic events among the population)

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:
**N/A
**Process measure

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 defined fields in electronic claims.
- No fees or licensure requirements are required.
- The developer notes that the measure has clear specifications but data methods and calculation methods may vary. Therefore, NCQA conducts an independent audit of all HEDIS collection and reporting processes as well as an audit of the data which are manipulated by those processes in order to verify that HEDIS specifications are met.

Questions for the Committee:
o *Are the required data elements routinely generated and used during care delivery?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3. Feasibility
Comments:
**All data elements are in fields defined by electronic claims submission
**Very feasible
**Data elements are clearly defined and available in electronic claims.

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**

Accountability program details:

- Medicaid Adult Core Set
- NCQA State of Health Care Quality Report
- NCQA Health Plan Ratings/Report Card
- NCQA Quality Compass
- NCQA Health Plan Accreditation

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

- The developer publicly reports rates across all plans and creates benchmarks to help plans how they perform compared to other plans.
- The developer publishes performance results and data annually in their Quality Compass tool and presents data at various conferences and webinars. The developer also provides regular technical assistance through its Policy Clarification Support System.
- The developer uses several methods to obtain input from users during its "reevaluation process," including, vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer noted that the health plans have not reported significant implementation barriers. Questions from users typically center around clarifications of the specifications such as benefit requirements and approved medications to identify the eligible population.

Additional Feedback:

- 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

- The developer notes that in the past 2 years, performance rates for this measure have been generally stable. In 2017, Medicaid plans had an average performance rate of 81 percent. The most significant variation is between the 10th and 90th percentiles, suggesting room for improvement.
- This measure was first introduced in HEDIS 2013. Rates for Medicaid were 78 percent. In the last 6 years, the developer has seen an improvement of 3 percent.

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 were reported by the developer.

Potential harms

- None were reported by the developer.

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:

**Current use in public reporting, Medicaid adult core set, quality compass, quality improvement annual state of healthcare quality, health care plan accreditation.

**It's unclear if the measure is actually moving the needle on improvement. Certainly, the knowledge of antipsychotic side effects has grown substantially.

**This measure is used by NCQA in 5 national reporting and accountability data sets. Feedback is achieved in a number of ways. Benchmarks are established to help plans with seeing how they perform compared to other plans. Developer uses feedback to provide TA to plans who request it.

4b1. Usability – Improvement

Comments:

**From 2015 to 2017, performance rates for this measure have been generally stable or shown slight improvement. In 2017, Medicaid plans had an average performance rate of 81 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement. In 2017, Medicaid plans in the 10th percentile had a rate of 74 percent, compared to 87 percent among plans in the 90th percentile. No negative unintended consequences have been identified.

**Generally benefits outweigh harms.

**As with other a number of other measures, improvement on this measure for the Medicaid population requires some special attention and likely incentives. No harms noted and benefits are considerable given the risks of diabetes for this population.

**The measures can lead to unnecessary testing and associated unnecessary pain, time and expense for patients.

Criterion 5: Related and Competing Measures

Related or competing measures

- 1933: Cardiovascular Monitoring for People with Cardiovascular Disease and Schizophrenia (SMC)
- 1934: Diabetes Monitoring for People with Diabetes and Schizophrenia (SMD)

Harmonization

- Specifications are harmonized to the extent possible, per the developer.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 1932
Measure Title: Diabetes Screening for People with Schizophrenia or Bipolar Disorder Who Are Prescribed Antipsychotic Medications (SSD)

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document
   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.

   The measure is specified and tested at the health plan level

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

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐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, skip Questions #3-8, then go to Question #9)

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

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

Reliability of the measure score was assessed using 2016 HEDIS data that included 202 Medicaid plans.

4. 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.*

REFERENCE: Testing attachment, section 2a2.2
***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 #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

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

REFERENCE: Testing attachment, section 2a2.2
***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 #6)

☐Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

The developer used a beta-binominal model to assess the signal-to-noise ratio. Results of reliability testing was 0.959

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

REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   REFERENCE: Testing attachment, section 2a2.2
   ***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 #8)
   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   ***TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify**.*
   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

***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 #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

Testing was not performed for exclusions.

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

This is a process measure

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

The developer compared performance between to randomly selected plans at the 25th and 75th percentile to understand the variation in performance. Using the t-test method, they calculated a testing statistic based on the sample size, performance rate, and standardized error of each plan, which was then compared against a normal distribution. The results showed that the two plans' performance was significantly different from each other.

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicaid	1,464	80.7	5.8	74.0	77.5	81.0	84.2	87.4	6.7	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

Measure not specified for more than one data source.

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)
No missing data

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***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 #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

To assess the validity of the measure, the developer conducted construct validity testing using the Pearson correlation coefficient to examine the association between using this measure and measure 1934, which both focus on patients with schizophrenia and whether they received care for diabetes. They found that there is a statistically significant (0.25) and positive relationship between the two measures.

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☐Yes (go to Question #22)
☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*
***REFERENCE:** Testing attachment, section 2b1.*
***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 #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?
NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***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 #25)
☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Measure Number (*if previously endorsed*)**:** 1932

Measure Title: Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)

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

Date of Submission: 4/2/2018

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: Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications

☐ 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.

Patients with schizophrenia or bipolar disorder>>increased risk for diabetes>>antipsychotic medications are an expected treatment and increase the risk of metabolic diseases>>screening for diabetes>>opportunity for early diagnosis and treatment, if warranted>>reduced poor health outcomes (e.g., premature mortality)

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

****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

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. APA Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition; 2004 Feb. 184 p. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf and GUIDELINE WATCH: PRACTICE GUIDELINE FOR THE TREATMENT OF PATIENTS WITH SCHIZOPHRENIA; 2009 SEP. 10 P. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia-watch.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.	**Acute Phase Treatment [A, A-, B, C, D, E, F, G]** • General medical health as well as medical conditions that could contribute to symptom exacerbation can be evaluated by medical history, physical and neurological examination, and appropriate laboratory, electrophysiological, and radiological assessments [I]. Measurement of body weight and vital signs (heart rate, blood pressure, temperature) is also recommended [II]. • Other laboratory tests to be considered to evaluate health status include a complete blood count (CBC); measurements of blood

	electrolytes, glucose, cholesterol, and triglycerides; tests of liver, renal, and thyroid function; a syphilis test; and when indicated and permissible, determination of HIV status and a test for hepatitis C [II]. **Stable Phase [A, A-, B, C, D, E, F, G]** • Routine monitoring for obesity-related health problems (e.g., high blood pressure, lipid abnormalities, and clinical symptoms of diabetes) and consideration of appropriate interventions are recommended particularly for patients with BMI in the overweight and obese ranges [II]. Clinicians may consider regular monitoring of fasting glucose or hemoglobin A1c levels to detect emerging diabetes, since patients often have multiple risk factors for diabetes, especially patients with obesity [I]
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline. Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are “blind” to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.

	[C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	[I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence.
Provide all other grades and definitions from the recommendation grading system	[III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	"Relevant literature was identified through a computerized search of PubMed for the period from 1994 to 2002. Using the keywords schizophrenia OR schizoaffective, a total of 20,009 citations were found. After limiting these references to clinical trials and meta-analyses published in English that included abstracts, 1,272 articles were screened by using title and abstract information. The Cochrane Database of Systematic Reviews was also searched by using the keyword schizophrenia. Additional, less formal literature searches were conducted by APA staff and individual members of the work group on schizophrenia. Sources of funding were considered when the work group reviewed the literature but are not identified in this document. When reading source articles referenced in this guideline, readers are advised to consider the sources of funding for the studies"

Estimates of benefit and consistency across studies	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline"
What harms were identified?	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	To our knowledge, there have been no published studies since the clinical practice guidelines that would contradict the current body of evidence

Table 2. ADA Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2018 Submission** American Diabetes Association (2018). Standards of medical care in diabetes--2018. Diabetes Care, 41, S28–S37. http://care.diabetesjournals.org/content/diacare/suppl/2017/12/08/41.Supplement_1.DC1/DC_41_S1_Combined.pdf **2012 Submission** American Diabetes Association (2011). Standards of medical care in diabetes--2011. Diabetes Care, 34, S11-61. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3006050/
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.	**2018 Submission** Annually screen people who are prescribed atypical antipsychotic medications for prediabetes or diabetes. (B) **2012 Submission** Testing to detect type 2 diabetes and assess risk for future diabetes in asymptomatic people should be considered in adults of any age who are overweight

	or obese (BMI greater than or equal to 25 kg/m2) and who have one or more additional risk factors for diabetes. Grade B Recommendation. 1. Testing should be considered in all adults who are overweight (BMI greater than or equal to 25 kg/m2*) and have additional risk factors: • physical inactivity • first-degree relative with diabetes • high-risk race/ethnicity (e.g., African American, Latino, Native American, Asian American, Pacific Islander) • women who delivered a baby weighing greater than 9 lb or were diagnosed with gestational diabetes mellitus • hypertension greater or equal to 140/90 mmHg or on therapy for hypertension) • HDL cholesterol level less than 35 mg/dl (0.90 mmol/l) and/or a triglyceride level greater than 250 mg/dl (2.82 mmol/l) • women with polycystic ovarian syndrome (PCOS) • A1c greater than or equal to 5.7%, IGT, or IFG on previous testing • other clinical conditions associated with insulin resistance (e.g., severe obesity, acanthosis nigricans) • history of CVD 2. In the absence of the above criteria, testing for diabetes should begin at age 45 years. 3. If results are normal, testing should be repeated at least at 3-year intervals, with consideration of more frequent testing depending on initial results and risk status.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2018 Submission** B: Supportive evidence from well-conducted cohort studies • Evidence from a well-conducted prospective cohort study or registry • Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study

	2012 Submission B: Supportive evidence from well-conducted cohort studies • Evidence from a well-conducted prospective cohort study or registry • Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study
Provide all other grades and definitions from the evidence grading system	**2018 Submission** **A:** Clear evidence from well-conducted, generalizable randomized controlled trials that are adequately powered, including • Evidence from a well-conducted multicenter trial • Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., "all or none" rule developed by the Centre for Evidence-Based Medicine at the University of Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including • Evidence from a well-conducted trial at one or more institutions • Evidence from a meta-analysis that incorporated quality ratings in the analysis **C:** Supportive evidence from poorly controlled or uncontrolled studies • Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results • Evidence from observational studies with high potential for bias (such as case series with comparison with historical controls) • Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation **E:** Expert consensus or clinical experience **2012 Submission**

	A: Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted multicenter trial • Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., "all or none" rule developed by Center for Evidence Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted trial at one or more institutions • Evidence from a meta-analysis that incorporated quality ratings in the analysis **C:** Supportive evidence from poorly controlled or uncontrolled studies • Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results • Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls) • Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation **E:** Expert consensus or clinical experience
Grade assigned to the **recommendation** with definition of the grade	**2018 Submission** No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. **2012 Submission** N/A
Provide all other grades and definitions from the recommendation grading system	**2018 Submission** No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. **2012 Submission**

	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2018 Submission** The ADA does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the ADA systematic review, we provide information on an additional systematic review that supports the ADA's recommendations in Table 2. **2012 Submission** 6; This measure is supported by evidence that suggests individuals with schizophrenia and bipolar disorder are at higher risk for diabetes than the general population and that use of certain antipsychotic medications increases this risk.
Estimates of benefit and consistency across studies	**2018 Submission** See Table 2. **2012 Submission** • Benefit: screening allows for an appropriate treatment to be administered, if warranted • Harms: potential false positives resulting from screening • Cost: the screening exam • The studies consistently show that individuals with schizophrenia and bipolar disorder are at higher risk for diabetes than the general population and that use of certain antipsychotic medications increases this risk.
What harms were identified?	**2018 Submission** See Table 2. **2012 Submission** Potential false positives resulting from screening
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2018 Submission** To our knowledge, there have been no published studies since the clinical practice guidelines that would impact the recommendations. **2012 Submission**

	N/A

Table 3. Systematic Review Supporting Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. World Psychiatry. 2016;15(2):166-174. https://doi.org/10.1002/wps.20309
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	"This meta-analysis aimed: a) to describe pooled frequencies of T2DM in people with SMI; b) to analyze the influence of demographic, illness and treatment variables as well as T2DM assessment methods (i.e., blood testing, self-report, charts); and c) to describe T2DM prevalence in studies directly comparing persons with each specific SMI diagnosis to general population samples." "T2DM prevalences were consistently elevated for each of the three diagnostic subgroups compared to the general population, and comparative meta-analyses found no significant differences across schizophrenia, schizophrenia spectrum disorders, bipolar disorder and MDD. Thus, other diagnostic-independent factors likely influence T2DM frequency, including hyperglycaemia following psychotropic medication use and long-term exposure to unhealthy lifestyle behaviors, as well as potential genetic factors linking psychiatric and medical risk…Patient self-report yielded numerically the lowest T2DM prevalences; the T2DM prevalence was significantly lower compared with chart review data. This finding is likely due to the fact that, in chart review studies, patients were followed back a longer time, extending the detection period. In line with this interpretation, there was a trend for retrospective studies to be associated with higher T2DM prevalences than prospective ones As there are differences in T2DM prevalences across assessment methods, it is recommended that fasting blood glucose measurements (ideally even oral glucose tolerance testing as the gold standard)

	should be obtained prior to the first prescription of antipsychotic medication. The frequency of glucose metabolism testing will depend on the patient's medical history and the prevalence of baseline risk factors. For patients on antipsychotic medication with normal baseline tests, it is recommended that measurements should be repeated at 12 weeks after initiation of treatment and at least annually thereafter, with more frequent assessments in high-risk patients, such as those with significant weight gain, post-partum diabetes or a first-degree family history of diabetes."
Grade assigned for the quality of the quoted evidence with definition of the grade	This systematic review was conducted in accordance with the M eta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614) and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097)
Provide all other grades and definitions of the evidence in the grading system	This systematic review was conducted in accordance with the M eta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614) and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097)
What is the time period covered by the body of evidence?	Database inception to August 1, 2015
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: 118 Quality of studies: "observational studies (cross-sectional, retrospective and prospective studies) and randomized controlled trials in adults with a psychiatric diagnosis of schizophrenia or related psychotic disorders, bipolar disorder or MDD according to the DSM-IV-TR or the ICD-10, irrespective of clinical setting (inpatient, outpatient or mixed, community setting), that reported study-defined T2DM prevalences."
What is the overall quality of evidence across studies in the body of evidence?	Overall, the quality of evidence supporting this measure is strong. There are over 100 studies in the

	evidence review that examine the prevalence and effectiveness of diabetes screening and monitoring for individuals with SMI, including schizophrenia and bipolar disorder. Further, the quality of studies included in the systematic review were well-designed observational studies and randomized control trials.
Estimates of benefit and consistency across studies in body of evidence– what are the estimates of benefits?	"To our knowledge, this is the first meta-analysis of T2DM including and comparing data from the three main SMIs, namely schizophrenia and related psychotic disorders, bipolar disorder and MDD. Approximately one in 10 individuals with SMI (11.3%; 95% CI: 10.0%-12.6%) had T2DM, and the relative risk for T2DM in multi-episode persons with SMI was almost double (RR=1.85, 95% CI: 1.45-2.37) that found in matched general population comparison samples. Our meta-analysis highlighted geographical differences in T2DM, mirroring the different prevalences in the general population, indicating the possible influence of lifestyle and other environmental factors with or without genetic risk differences. Thus, considering the observed increased T2DM risks, screening for and trying to minimize risk factors (including adverse lifestyle factors and specific antipsychotic medication choice) should be a key priority in the multidisciplinary treatment of people with SMI36-39. Our data clearly demonstrate that people with SMI should be considered as a "homogeneous and important high-risk group" that needs proactive screening for T2DM. There were no significant differences between the various treatment settings, and data collection before versus after the year 2000. There was also no difference in T2DM prevalence between population based and non-population based studies. In contrast, a higher T2DM prevalence was observed in studies relying upon clinical data gleaned from file and chart reviews versus self-report studies. A trend for higher T2DM was found in retrospective studies versus cross-sectional (p=0.054) and versus prospective (p=0.053) studies."

What harms were studied and how to they affect the net benefit (benefits over harm)?	No harms associated with testing were identified in the evidence reviewed.

Table 4. Systematic Review

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2012 Submission** Marder, S. R., Essock, S. M., Miller, A. L., Buchanan, R. W., Casey, D. E., Davis, J. M., et al. (2004). Physical health monitoring of patients with schizophrenia. Am J Psychiatry, 161, 1334-1349. https://www.ncbi.nlm.nih.gov/pubmed/15285957
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.	**2012 Submission** Patients who have significant risk factors for diabetes (family history, BMI greater than or equal to 25, waist circumference greater than or equal to 35 inches for women and greater than or equal to 40 inches for men) should have their fasting plasma glucose level or hemoglobin A1c value monitored 4 months after starting an antipsychotic and then yearly.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2012 Submission** Level 2 evidence: data from cohort studies, outcomes research, or low-quality randomized, controlled studies
Provide all other grades and definitions from the evidence grading system	**2012 Submission** Clear evidence from multiple randomized, controlled trials was considered level-1 evidence; and data from case-control studies were considered level-3 evidence
Grade assigned to the **recommendation** with definition of the grade	**2012 Submission** Expert consensus with evidence review
Provide all other grades and definitions from the recommendation grading system	**2012 Submission** N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2012 Submission** 6; cohort studies, outcomes research, or low-quality randomized, control studies
Estimates of benefit and consistency across studies	**2012 Submission**

	The studies consistently show that individuals with schizophrenia and bipolar disorder are at higher risk for diabetes than the general population and that use of certain antipsychotic medications increases this risk.
What harms were identified?	**2012 Submission** Potential false positives resulting from screening
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2012 Submission** N/A

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 APA 2009 Guideline Watch identified a number of controlled clinical trials examining treatments to prevent or treat weight gain and metabolic changes caused by antipsychotic use. The Guideline Watch additionally cite several randomized control trials (RCTs) related to new antipsychotics used to treat schizophrenia. This report highlights research studies published since the 2004 APA Practice Guidelines for the Treatment of Patients with Schizophrenia and furthers the known link between metabolic side effects and antipsychotics used to treat schizophrenia.

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

"This watch highlights key research studies published since that date. The studies were identified by a MEDLINE literature search for meta-analyses and randomized, controlled trials published between 2002 and 2008, using the same key words used for the literature search performed for the 2004 guideline."

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

GUIDELINE WATCH: PRACTICE GUIDELINE FOR THE TREATMENT OF PATIENTS WITH SCHIZOPHRENIA; American Psychiatric Association, 2009 SEP. 10 P.

https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia-watch.pdf

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 1932
Corresponding Measures:
De.2. **Measure Title:** Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients 18 – 64 years of age with schizophrenia or bipolar disorder, who were dispensed an antipsychotic medication and had a diabetes screening test during the measurement year.
1b.1. **Developer Rationale:** As patients with schizophrenia or bipolar disorder are at an increased risk for diabetes, and antipsychotic medications are an expected treatment that increases the risk of metabolic diseases, screening for diabetes will allow for proper diagnosis and treatment, if warranted.

S.4. **Numerator Statement:** Among patients 18-64 years old with schizophrenia or bipolar disorder, those who were dispensed an antipsychotic medication and had a diabetes screening testing during the measurement year.
S.6. **Denominator Statement:** Patients ages 18 to 64 years of age as of the end of the measurement year (e.g., December 31) with a schizophrenia or bipolar disorder diagnosis and who were prescribed an antipsychotic medication.
S.8. **Denominator Exclusions:** Exclude members who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients with diabetes during the measurement year or the year prior to the measurement year.

Exclude patients who had no antipsychotic medications dispensed during the measurement year.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Health Plan, Integrated Delivery System, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable.

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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
1932_SSD_MEF_7.1.docx

1a.1 <u>For Maintenance of Endorsement:</u> 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.
Yes

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)*
<u>If a COMPOSITE</u> (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.
As patients with schizophrenia or bipolar disorder are at an increased risk for diabetes, and antipsychotic medications are an expected treatment that increases the risk of metabolic diseases, screening for diabetes will allow for proper diagnosis and treatment, if warranted.

1b.2. **Provide performance scores on the measure as specified (<u>current and over time</u>) at the specified level of analysis**. *(<u>This is required for maintenance of endorsement</u>. 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 following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are stratified by year and product line (i.e. Medicaid).

Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications– Medicaid Rate (HMO and PPO Combined)
MEASUREMENT YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 79.8% | 0.1 | 72.7% | 75.7% | 80.1% | 83.8% | 87.0% | 8.1
2016 | 80.4% | 0.1 | 72.3% | 77.4% | 80.7% | 84.0% | 87.2% | 6.6
2017 | 80.7% | 0.1 | 74.0% | 77.5% | 81.0% | 84.2% | 87.4% | 6.7

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 47 million Medicaid health plan beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the mean eligible population for the measure across health plans.

Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications– Medicaid
YEAR | N Plans | Median Denominator Size per plan
2015 | 143 | 437
2016 | 185 | 804
2017 | 202 | 1,018

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.** *(<u>This is required for maintenance of endorsement</u>. 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.*

HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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

A number of research studies, including several meta-analyses, demonstrate that individuals with serious mental illness have an increased risk for diabetes as well as disparities in their care.

One review article estimated the prevalence of diabetes among individuals with SMI is approximately 12% (Holt and Mitchell, 2015), while the prevalence in the general population is approximately 9% aged =18 (CDC, 2017). Additionally, there is a known link between SMI treatments such as mood stabilizers, anticonvulsants and antipsychotic medications to adverse metabolic risks in patients, such as diabetes (Vancampfort, 2016).

A systematic review article assessed 118 cross-sectional, retrospective and prospective studies, and population versus non-population based studies comparing SMI individuals with non-serious mental illness control groups. Based on this evidence review, authors conclude that diabetes is more common among patients with SMI with a relative risk of 2.04 in patients with schizophrenia or related psychotic disorders and 1.89 in patients with bipolar disorder compared to the general population (Vancampfort, 2016).

Evidence suggests that individuals with SMI, specifically those with schizophrenia and bipolar disorder, are at increased risk of developing diabetes due to a higher prevalence of risk factors including tobacco use, poor nutrition and obesity and weight gain from the use of antipsychotics (Mangurian, 2016). Furthermore, these risk factors result in increased morbidity, such as hospitalizations and complications from diabetes, and mortality in the SMI population (Mai et al., 2011; CDC, 2010).

Despite these risks, people with SMI are less likely to have annual A1c testing or glucose screening (Banta 2009; Mai, 2011; Mangurian et al., 2016). A literature review found that up to 70% of individuals on antipsychotics do not receive screening or treatment for diabetes (Mangurian et al., 2016). In another study, only 47.3% of Medicaid psychiatric patients received annual HbA1c testing. Further, researchers in this study found that second-generation antipsychotic medications, used for schizophrenic and bipolar patients, were associated with higher diabetes risk and a reduced likelihood of HbA1c testing. (Banta, 2009)

Another study found that only 37.2% of mental health patients, compared to 42.9% of non-mental health patients, received a recommended HbA1c annual test. In general, patients with mental illness received less ongoing diabetes monitoring and had higher risk for diabetes complications and diabetes-related mortality compared to non-mental health patients (Mai, 2011).

References

Banta JE, Morrato EH, Lee SW, et al. (2009) Retrospective Analysis of Diabetes Care in California Medicaid Patients with Mental Illness. J Gen Intern Med. 24:802-8.

Centers for Disease Control and Prevention (CDC). (2010) Diagnosed and undiagnosed diabetes in the United States, all ages, 2010. Retrieved from: http://www.cdc.gov/diabetes/pubs/estimates11.htm. Accessed on June 19, 2014.

Centers for Disease Control and Prevention (CDC). National Diabetes Statistics Report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2017.

Holt R.I., Mitchell A.J. (2015). Diabetes mellitus and severe mental illness: mechanisms and clinical implications. Nat Rev Endocrinol. 2015 Feb;11(2):79-89. doi: 10.1038/nrendo.2014.203.

Mai Q, Holman CD, Sanfilippo FM, et al. (2011) Mental illness related disparities in diabetes prevalence, quality of care and outcomes: a population-based longitudinal study. BMC Medicine. 9:118.

Mangurian, C., Newcomer, J.W., Modlin, C. et al. Diabetes and Cardiovascular Care Among People with Severe Mental Illness: A Literature Review. J GEN INTERN MED (2016) 31: 1083. https://doi.org/10.1007/s11606-016-3712-4

Mangurian C, Newcomer JW, Vittinghoff E, Creasman JM, Knapp P, Fuentes-Afflick E, Schillinger D. Diabetes Screening Among Underserved Adults With Severe Mental Illness Who Take Antipsychotic Medications. JAMA Intern Med. 2015;175(12):1977–1979. doi:10.1001/jamainternmed.2015.6098

Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. World Psychiatry. 2016;15(2):166-174. doi:10.1002/wps.20309.

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):*
Behavioral Health, Endocrine : Diabetes

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

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

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 Applicable

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:** 1932_SSD_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.
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.
No important changes since the last update.

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).
Among patients 18-64 years old with schizophrenia or bipolar disorder, those who were dispensed an antipsychotic medication and had a diabetes screening testing during the measurement year.

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).
A glucose test (Glucose Tests Value Set) or an HbA1c test (HbA1c Tests Value Set) performed during the measurement year, as identified by claim/encounter or automated laboratory data.

See corresponding Excel document for the Glucose Tests Value Set and the HbA1c Tests Value Set.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients ages 18 to 64 years of age as of the end of the measurement year (e.g., December 31) with a schizophrenia or bipolar disorder diagnosis and who were prescribed an antipsychotic medication.

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).
Follow the steps below to identify the eligible population.

Identify members with schizophrenia or bipolar disorder as those who met at least one of the following criteria during the measurement year.

- At least one acute inpatient encounter, with any diagnosis of schizophrenia or bipolar disorder. Any of the following code combinations meet criteria:
  - BH Stand Alone Acute Inpatient Value Set with Schizophrenia Value Set.
  - BH Stand Alone Acute Inpatient Value Set with Bipolar Disorder Value Set.
  - BH Stand Alone Acute Inpatient Value Set with Other Bipolar Disorder Value Set.
  - BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set with Schizophrenia Value Set.
  - BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set with Bipolar Disorder Value Set.
  - BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set with Other Bipolar Disorder Value Set.
- At least two visits in an outpatient, intensive outpatient, partial hospitalization, ED or nonacute inpatient setting, on different dates of service, with any diagnosis of schizophrenia. Any two of the following code combinations meet criteria:
  - BH Stand Alone Outpatient/PH/IOP Value Set with Schizophrenia Value Set.
  - BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set with Schizophrenia Value Set.
  - ED Value Set with Schizophrenia Value Set.
  - BH ED Value Set with ED POS Value Set with Schizophrenia Value Set.
  - BH Stand Alone Nonacute Inpatient Value Set with Schizophrenia Value Set.
  - BH Nonacute Inpatient Value Set with BH Nonacute Inpatient POS Value Set with Schizophrenia Value Set.
- At least two visits in an outpatient, intensive outpatient, partial hospitalization, ED or nonacute inpatient setting, on different dates of service, with any diagnosis of bipolar disorder. Any two of the following code combinations meet criteria:
  - BH Stand Alone Outpatient/PH/IOP Value Set with Bipolar Disorder Value Set.
  - BH Stand Alone Outpatient/PH/IOP Value Set with Other Bipolar Disorder Value Set.
  - BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set with Bipolar Disorder Value Set.

- BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set with Other Bipolar Disorder Value Set.
- ED Value Set with Bipolar Disorder Value Set.
- ED Value Set with Other Bipolar Disorder Value Set.
- BH ED Value Set with ED POS Value Set with Bipolar Disorder Value Set.
- BH ED Value Set with ED POS Value Set with Other Bipolar Disorder Value Set.
- BH Stand Alone Nonacute Inpatient Value Set with Bipolar Disorder Value Set.
- BH Stand Alone Nonacute Inpatient Value Set with Other Bipolar Disorder Value Set.
- BH Nonacute Inpatient Value Set with BH Nonacute Inpatient POS Value Set with Bipolar Disorder Value Set.
- BH Nonacute Inpatient Value Set with BH Nonacute Inpatient POS Value Set with Other Bipolar Disorder Value Set.

(See corresponding Excel document for the above value sets)

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude members who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients with diabetes during the measurement year or the year prior to the measurement year.

Exclude patients who had no antipsychotic medications dispensed during the measurement year.

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.)*
Exclude members who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These members may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

Patients are excluded from the denominator if they have diabetes (during the measurement year or the year prior to the measurement year). There are two ways to identify patients with diabetes: 1) pharmacy data or 2) claim/encounter data. Both methods should be used to identify patients with diabetes, but a patient only needs to be identified by one method to be excluded from the measure. Members may be identified as having diabetes during the measurement year or the year prior to the measurement year.

Pharmacy data: Patients who were dispensed insulin or oral hypoglycemics/antihyperglycemics during the measurement year or year prior to the measurement year on an ambulatory basis (Diabetes Medications List).

Claim/encounter data: Patients who met at any of the following criteria during the measurement year or the year prior to the measurement year (count services that occur over both years).
- At least two outpatient visits (Outpatient Value Set), observation visits (Observation Value Set), ED visits (ED Value Set) or nonacute inpatient encounters (Nonacute Inpatient Value Set) on different dates of service, with a diagnosis of diabetes (Diabetes Value Set). Visit type need not be the same for the two encounters.
- At least one acute inpatient encounter (Acute Inpatient Value Set) with a diagnosis of diabetes (Diabetes Value Set).

PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES (Diabetes Medications List):
Alpha-glucosidase inhibitors:
Acarbose, Miglitol

Amylin analogs:
Pramlinitide

Antidiabetic combinations:
Alogliptin-metformin, Alogliptin-pioglitazone, Canagliflozin-metformin, Dapagliflozin-metformin, Empaglifozin-linagliptin, Empagliflozin-metformin, Glimepiride-pioglitazone, Glimepiride-rosiglitazone, Glipizide-metformin, Glyburide-metformin, Linagliptin-metformin, Metformin-pioglitazone, Metformin-repaglinide, Metformin-rosiglitazone, Metformin-saxagliptin, Metformin-sitagliptin, Sitagliptin-simvastatin

Insulin:

Insulin aspart, Insulin aspart-insulin aspart protamine, Insulin degludec, Insulin detemir, Insulin glargine, Insulin glulisine, Insulin isophane human, Insulin isophane-insulin regular, Insulin lispro, Insulin lispro-insulin lispro protamine, Insulin regular human, Insulin human inhaled

Meglitinides:
Nateglinide, Repaglinide

Glucagon-like peptide-1 (GLP1) agonists:
Dulaglutide, Exenatide, Liraglutide, Albiglutide

Sodium glucose cotransporter 2 (SGLT2) inhibitor:
Canagliflozin, Dapagliflozin, Empagliflozin

Sulfonylureas:
Chlorpropamide, Glimepiride, Glipizide, Glyburide, Tolazamide, Tolbutamide

Thiazolidinediones:
Pioglitazone, Rosiglitazone

Dipeptidyl peptidase-4 (DDP-4) inhibitors:
Alogliptin, Linagliptin, Saxagliptin, Sitaglipin

Exclude patients who had no antipsychotic medications dispensed during the measurement year. There are two ways to identify dispensing events: by claim/encounter data and by pharmacy data. The organization must use both methods to identify dispensing events, but an event need only be identified by one method to be counted.
- Claim/encounter data. An antipsychotic medication (Long-Acting Injections Value Set).
- Pharmacy data. Dispensed an antipsychotic medication (Antipsychotic Medications List; Antipsychotic Combination Medications List) on an ambulatory basis.

ANTIPSYCHOTIC MEDICATIONS:
(Antipsychotic Medications List)
Miscellaneous antipsychotic agents:
Aripiprazole, Asenapine, Brexpiprazole, Cariprazine, Clozapine, Haloperidol, Iloperidone, Loxapine, Lurisadone, Molindone, Olanzapine, Paliperidone, Pimozide, Quetiapine, Quetiapine fumarate, Risperidone, Ziprasidone

Phenothiazine antipsychotics:
Chlorpromazine, Fluphenazine, Perphenazine, Prochlorperazine, Thioridazine, Trifluoperazine

Thioxanthenes:
Thiothixene

Long-acting injections:
Aripiprazole, Fluphenazine decanoate, Haloperidol decanoate, Olanzapine, Paliperidone palmitate, Risperidone

(Antipsychotic Combination Medications List)
Psychotherapeutic combinations:
Fluoxetine-olanzapine, Perphenazine-amitriptyline

See corresponding Excel document for the value sets referenced above.

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.)*
None.

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.)*
Step1. Determine the eligible population: identify patients 18-64 years of age by the end of the measurement year.
Step 2. Search for an exclusion in the patient's history: Exclude patients from the eligible population if they meet the following criteria:
- Patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.
- Patients with diabetes during the measurement year or the year prior to the measurement year.
- Patients who had no antipsychotic medications dispensed during the measurement year.

Step 3. Determine the numerator: the number of patients who had a diabetes screening test during the measurement year.
Step 4. Calculate the 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.

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

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.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via NCQA's online data submission system.

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)*
Health Plan, Integrated Delivery System, Population : Regional and State

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Other, Outpatient Services
If other: Any outpatient setting represented with Medicaid claims data

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
1932_-_SSD_-_Testing_Form_v7.1_FINAL.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** 1932
Measure Title: Diabetes screening for people with schizophrenia or bipolar disorder who are prescribed antipsychotic medications (SSD)
Date of Submission: 4/2/2018
Type of Measure:

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

Instructions

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

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

2018 Submission

N/A

2012 Submission

N/A

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

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

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

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

2018 Submission

Population for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 202 Medicaid plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Population for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 145 Medicaid health plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

2012 Submission

Using Medicaid Analytic Extract (MAX) claims data from 2007 we included Medicaid recipients from 22 states who met the following criteria (1) enrolled in fee-for-service plans* (2) disability as the basis of eligibility; and (3) continuously enrolled in Medicaid for 10 months.
The data came from the following states: Alabama, Alaska, California, Connecticut, DC, Georgia, Idaho, Illinois, Indiana, Iowa, Louisiana, Maryland, Missouri, Mississippi, Nevada, New Hampshire, North Carolina, North Dakota, Oklahoma, South Dakota, 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*)

2018 Submission

Patient sample for measure score reliability testing: In 2016, HEDIS measures covered 47 million Medicaid members. Data are summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of Plans	Median number of eligible patients per plan
Medicaid	202	1,018

Beneficiary Sample for Construct Validity Testing: In 2016, HEDIS measures covered 47 million Medicaid beneficiaries. Data is summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of plans	Median number of eligible patients per plan
Medicaid	202	1,018

2012 Submission
We drew two analytic samples from the beneficiaries. Beneficiaries who had a primary diagnosis of schizophrenia on either one inpatient or two outpatient claims on different days were included in our schizophrenia sample. We also tested beneficiaries who had a primary diagnosis of either schizophrenia or bipolar disorder on either one inpatient or two outpatient claims on different days. Overall, there were 98,412 beneficiaries in the schizophrenia sample and 130,529 beneficiaries in the schizophrenia or bipolar disorder sample.

Beneficiaries ranged in age from 25 – 64 years. Just under half of the schizophrenia population was female (49.2%) while nearly 55% of beneficiaries with schizophrenia or bipolar disorder were female (54.8%). About 7% of both samples were Hispanic and African-Americans comprised 34% and 39%, respectively, of the schizophrenia or bipolar disorder and schizophrenia samples.
(*Beneficiaries enrolled in managed care plans (e.g. BHO or HMO plans) that provided usable claims records were included. About 1% of the schizophrenia sample was enrolled in a BHO (1.4%) and 11.5% were enrolled in an HMO).

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
2018 Submission
N/A

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
2018 Submission
We did not analyze performance by social risk factors.

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
2018 Submission
Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta.

Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

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

2012 Submission

The relevant unit of analysis for the proposed measures is aggregated state-level performance. Therefore, we conducted an analysis of test-retest reliability for state results to assess the reliability of state-level performance. To assess stability of state-level performance over time, we computed quartiles of performance based on the state distribution for each measure and assigned each state a score reflecting each state's performance relative to other states in the distribution during the measurement year. For example, a state in the top quartile of all states in 2007 for a given measure would be assigned a performance quartile score of '1' for 2007. This method was replicated for each measure. Next, we repeated this method using 2008 claims data and examined stability of performance quartile between 2007 and 2008.

We also report Pearson correlations measuring the association between 2007 and 2008 measure performance for the 16 states with data.

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

2018 Submission

Beta-Binomial Statistic:

Medicaid
0.959

2012 Submission

Overall, 4 of 16 states (25%) had no change in performance quartile between 2007 and 2008. State performance for this measure correlated at r=0.33. In general, the measure showed good test-retest reliability. The result also indicated that 2007 performance on this measure accounted for 11% of the variance in 2008 scores.

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?*)

Interpretation of measure score reliability testing: The testing suggests the measure has strong reliability with beta binomial result of 0.959 exceeding the 0.7 threshold.

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

2018 Submission

We assessed construct and face validity for this measure.

Method of testing construct validity: We tested for construct validity by exploring whether the Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications measure is correlated with the Diabetes Monitoring for People With Diabetes and Schizophrenia measure. We hypothesized that organizations that perform well on Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications should perform well on the Diabetes Monitoring for People With Diabetes and Schizophrenia measure because the two measures both focus on patients with schizophrenia and whether they received care for diabetes.

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable

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

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

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

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. On average, NCQA receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients,

policy makers and advocates. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQA's Board of Directors will be included in the next HEDIS year and reported as first-year measures.

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

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

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.
Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year's HEDIS Volume 2.

2012 Submission
Validity was assessed using several complementary methods.

Face validity was assessed through a multistakeholder Technical Advisory Group responsible for overseeing measure development. Additionally, face validity was captured through a public comment period and a series of focus groups involving the Medicaid Medical Directors Learning Network, Managed Behavioral Health Care Organizations, and State Mental Health Commissioners and Medical Directors. The panelists assessed the usability and feasibility of the measures.

Concurrent validity was assessed via Medicaid resource utilization from the Medicaid claims data. We examined rates of schizophrenia-related hospital and emergency room utilization as well as total Medicaid costs comparing beneficiaries in the highest and lowest performance quartiles for each measure.

Convergent and discriminant validity were assessed using the Medicaid Analytic Extract (MAX) from Medicaid claims in using 2007 data. Pearson correlation coefficients were used to assess measure correlations. We hypothesized similar measures (e.g. screening and monitoring) would be correlated and (b) process measures

would have negative correlations with measures of adverse events (e.g. mental health emergency room utilization).

2b1.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
2018 Submission
Statistical results of construct validity testing: The results in Table 1 indicate that there is a statistically significant (P<0.05) and positive relationship between the Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications measure and the Diabetes Monitoring for People With Diabetes and Schizophrenia measure.

Table 1. Correlations in Medicaid Measures – 2016

	Pearson Correlation Coefficient
	Diabetes monitoring for people with diabetes and schizophrenia
Diabetes screening for people with schizophrenia or bipolar disorder who are using antipsychotic medications	0.25

Note: p<0.05

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

2012 Submission:
Face validity:
The measures were deemed important, usable, and feasible to collect by the Technical Advisory Group overseeing the measure development, as well as focus groups with the Medicaid Medical Directors Learning Network, Managed Behavioral Healthcare Organizations, and State Mental Health Commissioners and Medical Directors.

Among 22 states, the measure had a minimum value of 2.3%, mean=12.1%, 25th percentile=8.4%, median=10.3%, 75th percentile=16.7% and a maximum value of 28.2%.

Concurrent validity:
Beneficiaries in the lowest performing states for the measure had higher rates of schizophrenia related hospitalization and ED use (24.3% and 26.6%, respectively) than individuals in the highest performing states (18.1% and 24.5%, respectively).

Concurrent and discriminant validity:
Performance on the measure was significantly correlated with cardiovascular screening (r=.276, p<.001).

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

Interpretation of construct validity testing: The two measures had statistically significant positive correlation, which indicates the measure has good construct validity.

Interpretation of systematic assessment of face validity: NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Diabetes screening for people with*

schizophrenia or bipolar disorder who are prescribed antipsychotic medications (SSD) is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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*)

Testing was not performed for exclusions.

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

Testing was not performed for exclusions.

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

Testing was not performed for exclusions.

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

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

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)*

2018 Submission

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

2012 Submission
Pearson correlations, means and percentiles are reported.

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

2018 Submission
HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicaid	1,464	80.7	5.8	74.0	77.5	81.0	84.2	87.4	6.7	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

2012 Submission
Among 22 states, the measure had a minimum value of 2.3%, mean=12.1%, 25th percentile=8.4%, median=10.3%, 75th percentile=16.7% and a maximum value of 28.2%.

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

2018 Submission
The difference between the 25th and 75th percentile is statistically significant for the Medicaid product line. For Medicaid plans, there is a 6.7 percentage point gap between 25th and 75th percentile plans. This gap represents an average 98 more patients with schizophrenia or bipolar disorder who were dispensed an antipsychotic medication having diabetes screening test during the measurement year in high performing Medicaid plans compared to low performing plans (estimated from average health plan eligible population).

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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)
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.
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures

3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Specific Plan for Use	Current Use (for current use provide URL)
	Public Reporting Medicaid Adult Core Set https://www.medicaid.gov/medicaid/quality-of-care/downloads/performance-measurement/2018-adult-core-set.pdf Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Health Plan Ratings https://reportcards.ncqa.org/#/health-plans/list Regulatory and Accreditation Programs Accreditation http://www.ncqa.org/Programs/Accreditation/Health-Plan-HP.aspx Quality Improvement (external benchmarking to organizations) Annual State of Health Care Quality: http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality

	Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass

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

NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. In 2017, the report included results from calendar year 2016 for health plans covering over 171 million people.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states were included in the rankings.

MEDICAID ADULT CORE SET: The Affordable Care Act (Section 1139B) requires the Secretary of HHS to identify and publish a core set of health care quality measures for adult Medicaid enrollees. The law requires that measures designated for the core set be currently in use. CMS annually releases information on state progress in reporting the Adult Core Set measures and assesses state-specific performance for measures that are reported by at least 25 states and which met internal standards of data quality.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

NCQA HEALTH PLAN ACCREDITATION: This measure is used to calculate health plan ratings, which are reported on the NCQA website. These ratings are based on a plan's performance on their HEDIS, CAHPS and accreditation standards scores. In 2017, a total of 521 Medicare Advantage health plans, 614 commercial health plans and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.

NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System.

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.
NCQA measures are evaluated regularly. During this "reevaluation" process, we seek broad input on the measure, including input on performance and implementation experience. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have not reported significant barriers to implementing this measure, as it uses the administrative data collection method. Questions have generally centered around minor clarification of the specifications, such as benefit requirements to report the measure and approved medications to identify the eligible population. NCQA responded to all questions to ensure consistent implementation of the specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities, like the Medicaid Adult Core Set.

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.**
Feedback has not required modification to this measure.

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.
From 2015 to 2017, performance rates for this measure have been generally stable or shown slight improvement. In 2017, Medicaid plans had an average performance rate of 81 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement. In 2017, Medicaid plans in the 10th percentile had a rate of 74 percent, compared to 87 percent among plans in the 90th percentile.

This measure was first introduced in HEDIS 2013. Rates for Medicaid were 78.0 percent. In the last 5 years, we have seen improvement of three percent.

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 were no identified unintended consequences for this measure during testing or since implementation.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unintended consequences for this measure during testing or since implementation.

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)**
1933 : Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)
1934 : Diabetes Monitoring for People With Diabetes and Schizophrenia (SMD)

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

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.**
N/A

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.)
N/A

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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance

Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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 Technical Advisory Group advised Mathematica Policy Research, Inc. and the National Committee for Quality Assurance during measure development. The TAG was responsible for providing feedback on measure concepts, specifications, results from field and data testing. The TAG consisted of a multistakeholder group of experts with knowledge in behavioral health and quality measurement.

Technical Advisory Group Roster:
Alisa Busch, MD, MS
Enola Proctor, PhD, MSW
David Shern, PhD
Wilma Townsend, MSW
Dan Ford, MD, MPH
Lorrie Rickman-Jones, PhD
Eric Hamilton
Alexander Young, MD, MHS
Peter Delany, PhD
Ben Druss, MD, MPH
Maureen Corcoran, MSN, MBA
Mike Fitzpatrick, MSW
Anita Yuskauskas
Bob Heinssen, PhD

Consultants:
Lisa Dixon, MD, MPH
Julie Kreyenbul, PharmD, PhD

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers
Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

Behavioral Health Measurement Advisory Panel:
Katharine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute
Christopher Dennis, MD, MBA, FAPA, Landmark Health, LLC
Ben Druss, MD, MPH, Emory University
Frank Ghinassi, PhD, ABPP, Rutgers University Behavioral Health Care

Connie Horgan, ScD, Brandeis University
Laura Jacobus-Kantor, PhD, SAMHSA
Jeffrey Meyerhoff, MD, Optum
Harold Pincus, MD, College of Physicians and Surgeons, Columbia University, New York Presbyterian Hospital, RAND
Michael Schoenbaum, PhD, National Institute of Mental Health
John Straus, MD, Massachusetts Behavioral Health Partnership-A Beacon Health Options Company

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2012

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Every 3-5 years.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

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 "AS IS" WITHOUT WARRANTY OF ANY KIND.

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 1933
Measure Title: Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18 – 64 years of age with schizophrenia and cardiovascular disease, who had an LDL-C test during the measurement year.
Developer Rationale: Appropriate monitoring of individuals with schizophrenia and cardiovascular disease may lead to proper treatment and management, as necessary.

Numerator Statement: An LDL-C test performed during the measurement year.
Denominator Statement: Patients 18-64 years of age as of the end of the measurement year (e.g., December 31) with a diagnosis of schizophrenia and cardiovascular disease.
Denominator Exclusions: Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System, Population : Regional and State

Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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 monitoring patients diagnosed with cardiovascular disease and schizophrenia which may lead to reduced poor health outcomes and improved long-term clinical outcomes.
- The developer provides a systematic review of the evidence including:
    - American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition. **Recommendations within these guidelines range from Grade I (substantial clinical confidence) to Grade II (moderate clinical confidence).**
    - Ayerbe L, Forgnone I, Foguet-Boreu Q, et al. Disparities in the management of cardiovascular risk factors in patients with psychiatric disorders: a systematic review and meta-analysis (2018). **Included studies were all considered to be of good quality, with score ⩾8 in the 14-item quality checklist.**
    - Vancampfort, D, Stubbs, B, Mitchell, A.J, et al. Risk of metabolic syndrome and its components in people with schizophrenia and related psychotic disorders, bipolar disorder and major depressive disorder: a systematic review and meta-analysis. World (2015). **This systematic review was conducted in accordance with the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard.**
- In addition, the developer cites the APA 2009 Guideline Watch which cites additional RCTs and studies that have been completed since the 2004 APA Practice Guidelines for the Treatment of Patients with Schizophrenia.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided additional systematic reviews of evidence listed above.

Exception to evidence: N/A

Questions for the Committee:

- *The evidence provided by the developer is updated and directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *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: high; Consistency: high (Box 5) > Moderate (Box 5b) > High

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 summarized the performance data at the health plan level using HEDIS health plan performance rates from 2015-2017. The data is stratified by year and insurance type.

Cardiovascular Monitoring for People with Cardiovascular Disease and Schizophrenia (HMO & PPO Combined)

Measurement Year	# of Plans	Median Denom. Size per plan	Mean	St Dev	10th	25th	50th	75th	90th	Interquartile range

2015	34	152	76.2%	0.1	64.7%	70.0%	78.7%	83.3%	87.9%	13.3
2016	37	67	78.0%	0.1	63.3%	73.5%	80.0%	83.6%	88.4%	10.1
2017	53	72	77.5%	0.1	63.2%	72.7%	77.6%	84.6%	88.3%	11.9

- In the previous review of this measure (2012) the developer provided field tests results to show a performance gap. Among 22 states, the measure had a minimum value of 11.7%, mean=54.5%, 25th percentile=44.4%, median=59.6%, 75th percentile=67.3% and a maximum value of 85.7%.

Disparities

- The developer does not provide disparities data since HEDIS data is stratified by type of insurance. While not specified in this measure, this measure can also be stratified by demographic variables in order to assess the health care disparities.
- The developer provides a summary of research studies demonstrating that individuals with schizophrenia have an increased risk for cardiovascular disease as well as disparities in their care.

Questions for the Committee:

o *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:

**The face validity of the causal pathway is high. But again the direct association with patient oriented outcomes is lacking.
**Evidence to support significance of measure is strong.
**No concerns.
**Individuals with schizophrenia are at significantly elevated risk of cardiovascular disease. Further those diagnosed with CV conditions struggle to manage their CV symptoms because of sedentary lifestyle, tobacco use and side effects of antipsychotic medications. Evidence of early mortality for people with schizophrenia is overwhelming. Measuring monitoring of CV disease in this population is critically important. This is directly related to the goal of improving the health of these patients.
**Evidence well documented.

1b. Performance Gap

Comments:

**There is a gap, but there has been little or no change in that gap.
**Very little change in performance since 2013. For Medicaid, on average only 2% improvement since use of the measure? The developer argues that there is room for improvement because of wide range in 10th and 90th percentiles, but these are inherently outliers.
**Ongoing opportunity for improvement.
**Yes, the submission does include performance data. It definitely demonstrates a gap in care to warrant a national performance measure. The data from the HEDIS measures used by the Medicaid health plans did not stratify by subgroups.
**Gap in care supported by data. Medicaid plans had an average performance rate of 78 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement.

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.

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)?*
- *The 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 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**

Committee pre-evaluation comments

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

2a1. Reliability – Specifications

Comments:

**Clearly defined.

**Reliability is based solely on estimates from beta-binomial model because the unit of analysis is the health plan, given limitations of the NCQA data source. This is a HEDIS measure that is used for accreditation purposes by NCQA. Beta binomial statistic=.718

*The measure appears to be reliable based on the testing employed. There is no reason that this measure cannot be consistently implemented.

**Since cardiovascular disease is often not diagnosed in patients with schizophrenia, Why require a prior Dx of cardiovascular disease for the denominator. Why not have the denominator be all patients with schizophrenia received LDL-C test annually?

2a2. Reliability – Testing

Comments:

**No--not at the score level--> moderate.

**Usual limitations for HEDIS measures.

**No concerns.
**No.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:
**Moderate validity. I doubt there is any evidence that patients already being treated need a yearly LDL.
**They assessed construct validity by examining whether adherence rates to their Diabetes Monitoring measures for persons with diabetes and schizophrenia were similar to adherence rates to this using a Pearson correlation. Result: .66 (2016 data). They reported assessing face validity by describing the development and maintenance approach NCQA uses for their HEDIS measures. There were no results. They simply proclaim, "Input from our multi-stakeholder measurement advisory panels and those submitting to public comment indicate the measure has face validity."
**No concerns.
**The 2018 submission reveals testing with high validity. The analysis indicates comparable results -- particularly across CV disease and diabetes. There is no indication that missing data threatens the validity of the measure.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:
**As above. I again worry we give these measures a pass on the issue of disparities.
**No capacity for adjustment for social risk factors which is particularly relevant for this target population. Little capacity to stratify by demographics unless "the data are available to a plan"
**None of this is clear from the 2018 submission by the sponsor -- although it does reference the NCQA IQR calculator.

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 defined fields in electronic claims.
- No fees or licensure requirements are required.
- The developer notes that the measure has clear specifications but data methods and calculation methods may vary. Therefore, NCQA conducts an independent audit of all HEDIS collection and reporting processes as well as an audit of the data which are manipulated by those processes in order to verify that HEDIS specifications are met.

Questions for the Committee:
o *Are the required data elements routinely generated and used during care delivery?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3. Feasibility
Comments:
**Very feasible.
**Uses both admin claims and medical record abstraction which may be burdensome.
**No concerns/electronic claims data
**According to the submission all of the data is submitted in electronic form. It is not apparent that there are any data elements not produced during care delivery. According to the submission, NCQA conducts an independent audit of all HEDIS collection and reporting processes in order to verify that specifications are met.
**Feasibility could be increased if the positive dX for cardiovascular disease was not required for the denominator.

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**

Accountability program details:

- Physician Value-Based Payment Modifier
- NCQA State of Health Care Quality Report
- NCQA Health Plan Ratings/Report Card
- NCQA Quality Compass
- Physician Feedback/Quality and Resource Use Reports

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

- The developer publicly reports rates across all plans and creates benchmarks to help plans how they perform compared to other plans.
- The developer publishes performance results and data annually in their Quality Compass tool and presents data at various conferences and webinars. The developer also provides regular technical assistance through its Policy Clarification Support System.
- The developer uses several methods to obtain input from users during its "reevaluation process," including, vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer noted that the health plans have not reported significant implementation barriers. Questions from users typically center around clarifications of the specifications.

Additional Feedback:

- 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

- The developer notes that in the past 2 years, performance rates for this measure have been generally stable. In 2017, Medicaid plans had an average performance rate of 78 percent. The most significant variation is between the 10th and 90th percentiles, suggesting room for improvement.
- This measure was first introduced in HEDIS 2013. Rates for Medicaid were 67.8 percent. In the last 6 years, the developer has seen an improvement of 2 percent.

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 were reported by the developer

Potential harms

- None were reported by the developer.

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:

**OK. I am lukewarm that we are not looking at action taken rather than simply measurement, at this stage.
**No concerns.
**Yes, the submission that the measure is current used and publicly reported in the 4 separate systems, including the CMS Physician Value-Based Payment Modifier. Some of these systems are subject to intense feedback from providers and health plans.

4b1. Usability – Improvement

Comments:

**Benefits outweigh harms.
**Major concerns is little meaningful change in performance since measure used. The dx's for IVD's were not provided. The events for CVD (acute MI, CABG, PCI) seemed like distal outcomes related to adverse event and/or procedures that indicated that treatment for existing and likely more severe CVD were received. This measure does not address prevention or early intervention of high LDL.
**I think a fasting lipid profile would be more ideal.

**These performance results are critical to improving outcomes for individuals with schizophrenia and addressing early mortality in this population. The benefits of this measure far outweigh any possible unintended consequences.

Criterion 5: Related and Competing Measures

Related or competing measures

- 1932: Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
- 1934: Diabetes Monitoring for People with Diabetes and Schizophrenia (SMD)

Harmonization

- Specifications are harmonized to the extent possible, per the developer.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 1933

Measure Title: Cardiovascular Monitoring for People with Cardiovascular Disease and Schizophrenia (SMC)

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document

   ***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.

   The measure is specified at the health plan level

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

REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

☒Yes (go to Question #3)

☐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, skip Questions #3-8, then go to Question #9)

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

REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

☒Yes (go to Question #4)

☐No (skip Questions #4-5 and go to Question #6)

Reliability of the measure score was assessed using 2016 HEDIS data that included 53 Medicaid plans.

4. 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.*

REFERENCE: Testing attachment, section 2a2.2
***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 #5)

☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

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

REFERENCE: Testing attachment, section 2a2.2
***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 #6)

☒Moderate (go to Question #6)

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

☐Insufficient (go to Question #6)

The developer used a beta-binominal model to assess the signal-to-noise ratio. Results of reliability testing was 0.718

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

REFERENCE: Testing attachment, section 2a2.
***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   REFERENCE: Testing attachment, section 2a2.2
   ***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 #8)
   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.
   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

***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 #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

Testing was not performed for exclusions.

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

This is a process measure

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

The developer compared performance between to randomly selected plans at the 25th and 75th percentile to understand the variation in performance. Using the t-test method, they calculated a testing statistic based on the sample size, performance rate, and standardized error of each plan, which was then compared against a normal distribution. The results showed that the two plans' performance was significantly different from each other.

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicaid	72	77.5	9.9	63.2	72.7	77.6	84.6	88.3	11.9	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

Measure is not specified for more than one data source.

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

No missing data

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***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 #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☒Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

The developer assessed construct validity using a Pearson correlation coefficient to examine the association between using this measure and measure 1934, which both focus on patients with schizophrenia and whether their chronic condition (diabetes or cardiovascular disease) is being monitored. They found that there is a statistically significant (0.66) and positive relationship between the two measures.

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

REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*

☐Yes (go to Question #22)
☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

***REFERENCE:** Testing attachment, section 2b1.*
***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 #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***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 #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1932

Measure Title: Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)

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

Date of Submission: 4/2/2018

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: Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)

☐ 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.

Patients diagnosed with schizophrenia and cardiovascular disease>>health care provider monitors patient's cardiovascular>>proper treatment and management>>reduced poor health outcomes (e.g., premature mortality, serious complications of cardiovascular disease)>>improved long-term clinical outcomes (desired outcome)

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

****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

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. APA Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition; 2004 Feb. 184 p. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf and GUIDELINE WATCH: PRACTICE GUIDELINE FOR THE TREATMENT OF PATIENTS WITH SCHIZOPHRENIA; American Psychiatric Association, 2009 SEP. 10 P. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia-watch.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.	**Acute Phase Treatment [A, A-, B, C, D, E, F, G]** • General medical health as well as medical conditions that could contribute to symptom exacerbation can be evaluated by medical history, physical and neurological examination, and appropriate laboratory, electrophysiological, and radiological assessments [I]. Measurement of body weight and vital signs (heart rate, blood pressure, temperature) is also recommended [II]. • Other laboratory tests to be considered to evaluate health status include a complete blood count (CBC); measurements of blood electrolytes, glucose, cholesterol, and

	triglycerides; tests of liver, renal, and thyroid function; a syphilis test; and when indicated and permissible, determination of HIV status and a test for hepatitis C [II]. **Stable Phase [A, A-, B, C, D, E, F, G]** • Routine monitoring for obesity-related health problems (e.g., high blood pressure, lipid abnormalities, and clinical symptoms of diabetes) and consideration of appropriate interventions are recommended particularly for patients with BMI in the overweight and obese ranges [II]. Clinicians may consider regular monitoring of fasting glucose or hemoglobin A1c levels to detect emerging diabetes, since patients often have multiple risk factors for diabetes, especially patients with obesity [I]
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline. Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are “blind” to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.

	[C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	[I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence.
Provide all other grades and definitions from the recommendation grading system	[III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	"Relevant literature was identified through a computerized search of PubMed for the period from 1994 to 2002. Using the keywords schizophrenia OR schizoaffective, a total of 20,009 citations were found. After limiting these references to clinical trials and meta-analyses published in English that included abstracts, 1,272 articles were screened by using title and abstract information. The Cochrane Database of Systematic Reviews was also searched by using the keyword schizophrenia. Additional, less formal literature searches were conducted by APA staff and individual members of the work group on schizophrenia. Sources of funding were considered when the work group reviewed the literature but are not identified in this document. When reading source articles referenced in this guideline, readers are advised to consider the sources of funding for the studies"

Estimates of benefit and consistency across studies	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline"
What harms were identified?	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	To our knowledge, there have been no published studies since the clinical practice guidelines that would impact the recommendations.

Table 2. Systematic Review Supporting Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Ayerbe L, Forgnone I, Foguet-Boreu Q, González E, Addo J, Ayis S. Disparities in the management of cardiovascular risk factors in patients with psychiatric disorders: a systematic review and meta-analysis. Psychological Medicine, 2018; 1:1-9. doi: 10.1017/S0033291718000302 https://www.ncbi.nlm.nih.gov/pubmed/29490716
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	Prospective studies comparing rates of screening, diagnosis, treatment and control of cardiovascular risk factors (CVRFs) for individuals with and without psychotic disorders, including schizophrenia were reviewed. Meta-analyses were done to summarize the findings when possible.

	Studies found that patients with schizophrenia were less likely to have their blood pressure recorded and also used antihypertensive and lipid-lowering drugs less frequently than general populations.
Grade assigned for the quality of the quoted evidence with definition of the grade	Included studies were all considered to be of good quality, with score ⩾8 in the 14-item quality checklist (National Institute of Health 2016).
Provide all other grades and definitions of the evidence in the grading system	Additional grading was not provided.
What is the time period covered by the body of evidence?	Database inception to 25 January 2017
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity: 20 Quality: (1) Observational prospective studies reporting original research data and (2) Studies presenting differences in rates of screening, diagnosis, follow-up, treatment or control of hypertension or dyslipidemia, smoking habit, diabetes for patients with and without each of the following mental disorders: schizophrenia, depression, anxiety, bipolar or personality disorder, identified with a validated scale or clinical assessment.
What is the overall quality of evidence across studies in the body of evidence?	Overall, the quality of evidence supporting this measure is strong. There are 20 studies in the evidence review that examine the rates of screening, diagnosis, treatment and control of cardiovascular risk factors for individuals with psychiatric disorders, including schizophrenia. Further, the quality of studies included in the systematic review were well-designed observational studies and studies presenting disparities in care for patients with psychotic disorders.
Estimates of benefit and consistency across studies in body of evidence– what are the estimates of benefits?	"The risk of bias and overall methodological quality of the studies fitting the inclusion criteria was assessed using the Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies of the National Institute of Health (USA) (online Supplement 3) (National Institute of Health 2016). Studies were excluded if they were: (1) Conducted in specific patient sub-populations (e.g. patients receiving specific medication); (2) Interventional studies; (3) Only presented results of univariate analyses;

	(4) Using composite exposures (e.g. affective disorders) unless separate results for each of them were presented; (5) Exposure analysed as continuous variable (e.g. score in a depression scale instead of a medical diagnosis, or a validated score above a cut-off point, which are the methods for categorization commonly used in clinical practice (National Institute for Health & Care Excellence, 2009, 2011); (6) Exposure presented as syndromes or symptoms (e.g. psychosis or hallucinations) rather than distinct diagnoses, which are the categories from the commonly used by clinicians who manage CVRFs (World Health Organization, 1978, 2010; American psychiatric Association, 1994, 2013); (7) Reporting a composite outcome (e.g. metabolic syndrome) unless separate results for each of its component had been provided. The reason not to include composite outcomes is because, according to the guidelines, clinicians have to care for each and every CVRF, therefore understanding the disparities affecting the management of each individual one is clinically relevant (National Institute for Health & Care Excellence, 2016a, b; National Institute for Health & Care Excellence, 2017a, b, c).”
What harms were studied and how to they affect the net benefit (benefits over harm)?	No harms associated with testing were identified in the evidence reviewed.

Table 3. Systematic Review Supporting Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Vancampfort, D., Stubbs, B., Mitchell, A.J., De Hert, M., Wampers, M., Ward, P.B., Rosenbaum, S., Correll, C.U. Risk of metabolic syndrome and its components in people with schizophrenia and related psychotic disorders, bipolar disorder and major depressive disorder: a systematic review and meta-analysis. World Psychiatry, 2015; 14:3 (339-347). https://doi.org/10.1002/wps.20252
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	“The primary aim of this systematic review and meta-analysis was to assess the prevalence of Metabolic Syndrome (MetS) and its components in people with schizophrenia and related psychotic disorders, bipolar disorder and major depressive

	disorder, comparing subjects with different disorders and taking into account demographic variables and psychotropic medication use. The secondary aim was to compare the MetS prevalence in persons with any of the selected disorders versus matched general population controls." People with severe mental illness, including schizophrenia, have a 2-3 times higher risk for premature death than the general population. Cardiovascular disease is attributed to approximately 60% of the excess mortality among people with severe mental illness. A reduced likelihood to receive standard levels of medical care as well as obstacles in access to medical care heighten existing risk factors, including antipsychotic medication use and an unhealthy lifestyle. "People treated with all individual antipsychotic medications had a significantly ($p<0.001$) higher MetS risk compared to antipsychotic-naïve participants. MetS risk was significantly higher with clozapine and olanzapine (except vs. clozapine) than other antipsychotics, and significantly lower with aripiprazole than other antipsychotics (except vs. amisulpride). Compared with matched general population controls, people with severe mental illness had a significantly increased risk for MetS (RR = 1.58; 95% CI: 1.35-1.86; $p<0.001$) and all its components, except for hypertension ($p = 0.07$). These data suggest that the risk for MetS is similarly elevated in the diagnostic subgroups of severe mental illness. Routine screening and multidisciplinary management of medical and behavioral conditions is needed in these patients."
Grade assigned for the quality of the quoted evidence with definition of the grade	This systematic review was conducted in accordance with the M eta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614) and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097)

Provide all other grades and definitions of the evidence in the grading system	This systematic review was conducted in accordance with the M eta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614) and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097)
What is the time period covered by the body of evidence?	Database inception to January 1, 2015
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity: 198 Quality: "observational studies (cross-sectional, retrospective and prospective studies) in adults that fulfilled the following criteria: a) a diagnosis of schizophrenia or a related psychotic disorder, bipolar disorder or major depressive disorder according to the DSM-IV or ICD-10, irrespective of clinical setting (inpatient, outpatient or mixed); and b) a MetS diagnosis according to non-modified ATP-III, ATP-III-A, IDF or World Health Organization standards. For a randomized control trial, we extracted the variables of interest at baseline. There were no language or time restrictions."
What is the overall quality of evidence across studies in the body of evidence?	Overall, the quality of evidence supporting this measure is strong. There are almost 200 studies in the evidence review that examine the prevalence and effectiveness of cardiovascular disease monitoring for individuals with SMI, including schizophrenia. Further, the quality of studies included in the systematic review were well-designed randomized control trials and observational studies.
Estimates of benefit and consistency across studies in body of evidence– what are the estimates of benefits?	"Relative risk meta-analyses established that there was no significant difference in MetS in studies directly comparing schizophrenia (39.2%, 95% CI: 30.5%-48.3%; n = 2,338) versus bipolar disorder (35.5%, 95% CI: 27.0-44.3%; n = 2,077) (N = 10, RR = 0.92, 95% CI: 0.79%-1.06%; $\chi 2 = 1.33$, $p = 0.24$; $Q = 21.3$, $p<0.011$). Similarly, there were no differences in the study directly comparing bipolar disorder (29.2%, 95% CI: 14.5%-46.2%; n = 137) versus major depressive disorder (34.0%, 95% CI: 19.4%-50.3%; n = 176) (N = 4; RR = 0.87, 95% CI: 0.48- 1.55; $\chi 2 = 0.21$, $p = 0.64$; $Q = 7.73$, $p = 0.0518$). Only two studies directly compared MetS in people with schizophrenia and major depressive disorder, precluding meta-analytic

	calculations…MetS prevalences were consistently elevated for each of the three diagnostic subgroups compared to the general population, and comparative meta-analyses found no significant differences across schizophrenia, bipolar disorder and major depressive disorder."
What harms were studied and how to they affect the net benefit (benefits over harm)?	No harms associated with testing were identified in the evidence reviewed.

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.

2018 Submission

The APA 2009 Guideline Watch identified a number of controlled clinical trials examining treatments to prevent or treat weight gain and metabolic changes caused by antipsychotic use. The Guideline Watch additionally cite several randomized control trials (RCTs) related to new antipsychotics used to treat schizophrenia. This report highlights research studies published since the 2004 APA Practice Guidelines for the Treatment of Patients with Schizophrenia and furthers the known link between metabolic side effects and antipsychotics used to treat schizophrenia.

2012 Submission

Summary of Evidence of High Impact: Individuals with schizophrenia are more likely than the general population to have lifestyle risk factors for cardiovascular disease and mortality (Brown, 1997; Phelan, et al., 2001; McCreadie, 2003; Osborn, et al., 2006; de Leon & Diaz, 2005; Hennekens, et al., 2005). Evidence suggests a higher prevalence of cardiovascular disease, most particularly, in younger people with schizophrenia (Bresee et al., 2010). While some evidence suggests high non-treatment rates for hyperlipidemia in patients with schizophrenia (Nasrallah, et al., 2006), patients with schizophrenia and elevated blood cholesterol levels are 25% less likely to be prescribed statins compared to the general population (Redelmeier, et al., 1998). Cardiovascular health monitoring for individuals with schizophrenia may lead to proper treatment and control of blood lipid levels.

Directness of Evidence to the Specified Measure: The evidence suggests that individuals with schizophrenia have a higher prevalence of cardiovascular disease due to a variety of risk factors. Monitoring of cardiovascular health for individuals with schizophrenia will lead to proper treatment, if necessary.

Quality of Body of Evidence: This measure is supported by prevalence studies that suggest a higher rate of cardiovascular disease in individuals with schizophrenia.

Consistency of Results across Studies: There is consistent evidence that shows individuals with schizophrenia have a higher prevalence of cardiovascular disease than the general population.

Net Benefit: Benefit: Monitoring patients with cardiovascular disease and schizophrenia may allow for proper treatment, if warranted. Cost: The monitoring exam

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

2018 Submission

"This watch highlights key research studies published since that date. The studies were identified by a MEDLINE literature search for meta-analyses and randomized, controlled trials published between 2002 and 2008, using the same key words used for the literature search performed for the 2004 guideline."

2012 Submission

Selected individual studies (rather than entire body of evidence)

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

2018 Submission

GUIDELINE WATCH: PRACTICE GUIDELINE FOR THE TREATMENT OF PATIENTS WITH SCHIZOPHRENIA; American Psychiatric Association, 2009 SEP. 10 P.

https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia-watch.pdf

2012 Submission

Brown S. Excess mortality of schizophrenia: a meta-analysis. Br J Psychiatry. 1997;171:502-508.

Phelan, M., Stradins, L. & Morrison, S. (2001) Physical health of people with severe mental illness. BMJ, 322, 443– 444.

McCreadie, R. The Scottish Schizophrenia lifestyle group. (2003) Diet, smoking and cardiovascular risk in people with schizophrenia: descriptive study. British Journal of Psychiatry, 183, 534–539.

Osborn, D.J., King, M.B. & Nazareth, I. (2006) Risk for coronary heart disease in people with severe mental illness: cross-sectional comparative study in primary care. Br J Psychiatry, 188, 271–277

De Leon, J. & Diaz, F.J. (2005) A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. Schizophr Res, 76, 135-157.

Hennekens, C.H., Hennekens, A.R., Hollar, D., Casey, D.E. (2005). Schizophrenia and increased risks of cardiovascular disease. Am Heart J, 150, 1115-1121.

Bresee, L.C., Majumdar, S.R., Patten, S.B., Johnson, J.A. (2010). Prevalence of cardiovascular risk factors and disease in people with schizophrenia: a population-based study. Schizophr Res. 2010;117:75-82.

Nasrallah H.A., Meyer J.A., Goff DC., McEvoy J.P., Davis S.M., Stroup S., Lieberman J.A. (2006). Low rates of treatment for hypertension, dyslipidemia and diabetes in schizophrenia: Data from the CATIE schizophrenia trial sample at baseline. Schizophr Res, 86, 15-22.

Redelmeier, D.A., Siew, H.T., Booth, G.L. (1998) The treatment of unrelated disorders in patients with chronic medical diseases. N Engl J Med, 160, 313-21.

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 1933
Corresponding Measures:
De.2. **Measure Title:** Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients 18 – 64 years of age with schizophrenia and cardiovascular disease, who had an LDL-C test during the measurement year.
1b.1. **Developer Rationale:** Appropriate monitoring of individuals with schizophrenia and cardiovascular disease may lead to proper treatment and management, as necessary.

S.4. **Numerator Statement:** An LDL-C test performed during the measurement year.
S.6. **Denominator Statement:** Patients 18-64 years of age as of the end of the measurement year (e.g., December 31) with a diagnosis of schizophrenia and cardiovascular disease.
S.8. **Denominator Exclusions:** Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Health Plan, Integrated Delivery System, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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
1933_SMC_MEF_7.1_FINAL_update_4.11.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.
Yes

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.
Appropriate monitoring of individuals with schizophrenia and cardiovascular disease may lead to proper treatment and management, as necessary.

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 following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are stratified by year and product line (i.e. Medicaid).

Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC) (HMO and PPO combined)
MEASUREMENT YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 76.2% | 0.1 | 64.7% | 70.0% | 78.7% | 83.3% | 87.9% | 13.3
2016 | 78.0% | 0.1 | 63.3% | 73.5% | 80.0% | 83.6% | 88.4% | 10.1
2017 | 77.5% | 0.1 | 63.2% | 72.7% | 77.6% | 84.6% | 88.3% | 11.9

Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)
YEAR | N Plans | Median Denominator Size per plan
2015 | 34 | 152
2016 | 37 | 67
2017 | 53 | 72

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.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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**

A number of research studies, including several meta-analyses, demonstrate that individuals with schizophrenia have an increased risk for cardiovascular disease as well as disparities in their care.

One review article estimated the prevalence of cardiovascular disease among individuals with SMI is approximately 10% (Correll et al., 2017). A systematic review article assessed 198 cross-sectional, retrospective and prospective studies, and population versus non-population based studies comparing SMI individuals with non-serious mental illness control groups. Based on this evidence review, authors conclude that the prevalence of metabolic syndrome and its components, which are considered to be highly predictive of cardiovascular disease, is approximately 33% (Vancampfort et al., 2015), whereas the prevalence of cardiovascular disease in the general adult population is approximately 11% (CDC, 2016). Additionally, there is a known link between antipsychotic medications and adverse cardiac and metabolic effects (De Hert et al., 2012).

Evidence suggests that individuals with SMI, specifically those with schizophrenia, are at increased risk of developing metabolic syndrome and subsequent cardiometabolic disorders due to a higher prevalence of risk factors including poor diet, lack of physical activities, smoking, substance abuse, older age, higher body mass index and side effects from the use of antipsychotics (Ringen et al., 2014; Vancampfort et al., 2015). Furthermore, these risk factors result in higher incidences of morbidity and increased non-suicide related mortality in individuals with schizophrenia (Ringen et al., 2014; Olfson et al., 2015).

Despite these risks, people on antipsychotics, including individuals with schizophrenia, are less likely to receive routine, cardiovascular monitoring (Mitchell et al., 2011). One systematic review found that only 42% of individuals on antipsychotics had their cholesterol measured (Mitchell et al., 2011). In another review, the rate of lipid testing among individuals on antipsychotics was as low as 6% in certain study populations (Baller et al., 2015).

References

Baller, J.B., McGinty, E.E., Azrin, S.T. Juliano-Bult, D. and Daumit, G.L. (2015). Screening for cardiovascular risk factors in adults with serious mental illness: a review of the evidence. BMC Psychiatry, 15:55. https://doi.org/10.1186/s12888-015-0416-y

Blackwell DL, Villarroel MA. Tables of Summary Health Statistics for U.S. Adults: 2015 National Health Interview Survey. Centers for Disease Control and Prevention and the National Center for Health Statistics. 2016. Available from: http://www.cdc.gov/nchs/nhis/SHS/tables.htm.

Correll CU, Solmi M, Veronese N, et al. Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: a large-scale meta-analysis of 3,211,768 patients and 113,383,368 controls. World Psychiatry. 2017;16(2):163-180. doi:10.1002/wps.20420.

De Hert, M., Detraux, J., Van Winkel, R., Yu, W. & Correll, C.U., Nature Reviews Endocrinology volume 8, pages 114–126 (2012). doi:10.1038/nrendo.2011.156

Mitchell, A., Delaffon, V., Vancampfort, D., Correll, C., & De Hert, M. (2012). Guideline concordant monitoring of metabolic risk in people treated with antipsychotic medication: Systematic review and meta-analysis of screening practices. Psychological Medicine, 42(1), 125-147. doi:10.1017/S003329171100105X

Olfson M, Gerhard T, Huang C, Crystal S, Stroup TS. Premature Mortality Among Adults With Schizophrenia in the United States. JAMA Psychiatry. 2015;72(12):1172–1181. doi:10.1001/jamapsychiatry.2015.1737

Ringen PA, Engh JA, Birkenaes AB, Dieset I and Andreassen OA (2014) Increased mortality in schizophrenia due to cardiovascular disease – a non-systematic review of epidemiology, possible causes, and interventions. Front. Psychiatry 5:137. doi: 10.3389/fpsyt.2014.00137

Vancampfort, D. , Stubbs, B. , Mitchell, A. J., De Hert, M. , Wampers, M. , Ward, P. B., Rosenbaum, S. and Correll, C. U. (2015), Risk of metabolic syndrome and its components in people with schizophrenia and related psychotic disorders, bipolar disorder and major depressive disorder: a systematic review and meta-analysis. World Psychiatry, 14: 339-347. doi:10.1002/wps.20252

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):*
Behavioral Health, Cardiovascular, Cardiovascular : Hyperlipidemia

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

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Individuals with multiple chronic conditions

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 Applicable

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:** 1933_SMC_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.
Yes

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.
NCQA added a hospice exclusion to most HEDIS measures in 2016. The focus of hospice care is not to cure illnesses of patients, but rather to improve comfort and quality of life for those with less than six months to live. Most HEDIS quality measures are focused on health screenings or treatments that are not clinically appropriate or beneficial for those who are at end of life. Many of these screenings and treatments would also be uncomfortable for hospice patients, add undue burden and have no impact on improving length or quality of life. Therefore, including individuals who are receiving hospice in our HEDIS quality measures is inappropriate.

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).
An LDL-C test performed during the measurement year.

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).
An LDL-C test (LDL-C Tests Value Set) performed during the measurement year, as identified by claim/encounter or automated laboratory data.
- See corresponding Excel document for the LDL-C Tests Value Set

The organization may use a calculated or direct LDL.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18-64 years of age as of the end of the measurement year (e.g., December 31) with a diagnosis of schizophrenia and cardiovascular disease.

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).
Follow the steps below to identify the eligible population.
Step 1: Identify patients with schizophrenia as those who met at least one of the following criteria during the measurement year:
• At least one acute inpatient encounter with any diagnosis of schizophrenia. Either of the following code combinations meets criteria:
– BH Stand Alone Acute Inpatient Value Set with Schizophrenia Value Set.
– BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set with Schizophrenia Value Set.
• At least two visits in an outpatient, intensive outpatient, partial hospitalization, ED or nonacute inpatient setting, on different dates of service, with any diagnosis of schizophrenia. Any two of the following code combinations meet criteria:
– BH Stand Alone Outpatient/PH/IOP Value Set with Schizophrenia Value Set.
– BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set with Schizophrenia Value Set.
– ED Value Set with Schizophrenia Value Set.
– BH ED Value Set with ED POS Value Set with Schizophrenia Value Set.
– BH Stand Alone Nonacute Inpatient Value Set with Schizophrenia Value Set.
– BH Nonacute Inpatient Value Set with BH Nonacute Inpatient POS Value Set with Schizophrenia Value Set

Step 2: Identify patients from step 1 who also have cardiovascular disease. Members are identified as having cardiovascular disease in two ways: by event or by diagnosis. The organization must use both methods to identify the eligible population, but a patient need only be identified by one to be included in the measure.

Event. Any of the following during the year prior to the measurement year meet criteria:
• AMI. Discharged from an inpatient setting with an AMI (AMI Value Set). To identify discharges:
1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
2. Identify the discharge date for the stay.
• CABG. Members who had CABG (CABG Value Set) in any setting.
• PCI. Members who had PCI (PCI Value Set) in any setting (e.g., inpatient, outpatient, ED).

Diagnosis. Identify members with IVD as those who met at least either of the following criteria during both the measurement year and the year prior to the measurement year. Criteria need not be the same across both years.
• At least one outpatient visit (Outpatient Value Set) with a diagnosis of IVD (IVD Value Set).
• At least one acute inpatient encounter (Acute Inpatient Value Set) with a diagnosis of IVD (IVD Value Set).

(See corresponding Excel document for the above value sets)

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

Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

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.)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These patients may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

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. Determine the eligible population: identify patients 18-64 years of age by the end of the measurement year with a diagnosis of schizophrenia and cardiovascular disease
Step 2. Determine the numerator: the number of patients who had an LDL-C test during the measurement year
Step 3. Calculate the 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.
N/A

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

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.
This measure is based on administrative claims and medical record documentation collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via NCQA's online data submission system.

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)*
Health Plan, Integrated Delivery System, 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
1933_-SMC_-_Testing_Form_v7.1_FINAL.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing

Measure Number (*if previously endorsed*)**:** 1933
Measure Title: Cardiovascular monitoring for people with cardiovascular disease and schizophrenia (SMC)
Date of Submission: 1/5/2018
Type of Measure:

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

Instructions

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

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

2018 Submission

N/A

2012 Submission

N/A

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

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**

☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

2018 Submission

Population for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 53 Medicaid plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Population for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 53 Medicaid health plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

2012 Submission

Using Medicaid Analytic Extract (MAX) claims data from 2007 we included beneficiaries from 22 states who met the following criteria (1) enrolled in fee-for-service plans* (2) disability as the basis of eligibility; and (3) continuously enrolled in Medicaid for 10 months.

Data from the following states were included in analytic samples: Alabama, Alaska, California, Connecticut, DC, Georgia, Idaho, Illinois, Indiana, Iowa, Louisiana, Maryland, Missouri, Mississippi, Nevada, New Hampshire, North Carolina, North Dakota, Oklahoma, South Dakota, 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*)

2018 Submission

Patient sample for measure score reliability testing: In 2016, HEDIS measures covered 47 million Medicaid beneficiaries. Data are summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of Plans	Median number of eligible patients per plan
Medicaid	53	57

Beneficiary Sample for Construct Validity Testing: In 2016, HEDIS measures covered 47 million Medicaid beneficiaries. Data is summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of plans	Median number of eligible patients per plan
Medicaid	53	57

2012 Submission
From the beneficiaries, we drew two analytic samples. Beneficiaries who had a primary diagnosis of schizophrenia on either one inpatient or two outpatient claims on different days were included in our schizophrenia sample. Overall, there were 98,412 beneficiaries in the schizophrenia sample.

Beneficiaries ranged in age from 25 – 64 years. Just under half of the schizophrenia population was female (49.2%). About 7% and 34% of the sample was Hispanic and African-American, respectively.

(*Beneficiaries enrolled in managed care plans (e.g. BHO or HMO plans) that provided usable claims records were included. About 1% of the schizophrenia sample was enrolled in a BHO (1.4%) and 11.5% were enrolled in an HMO).

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
2018 Submission
N/A

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
2018 Submission
We did not analyze performance by social risk factors.

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
2018 Submission
Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the

confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

2012 Submission

The relevant unit of analysis for the proposed measures is aggregated state-level performance. Therefore, we conducted an analysis of test-retest reliability for state results to assess the reliability of state-level performance. To assess stability of state-level performance over time, we computed quartiles of performance based on the state distribution for each measure and assigned each state a score reflecting each state's performance relative to other states in the distribution during the measurement year. For example, a state in the top quartile of all states in 2007 for a given measure would be assigned a performance quartile score of '1' for 2007. This method was replicated for each measure. Next, we repeated this method using 2008 claims data and examined stability of performance quartile between 2007 and 2008.

We also report Pearson correlations measuring the association between 2007 and 2008 measure performance for the 16 states with data.

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

2018 Submission

Beta-Binomial Statistic:

Medicaid
0.718

2012 Submission

Overall, 5 of 16 states (31%) had no change in performance quartile between 2007 and 2008. State performance was correlated at r=0.40. In general, the measure showed good test-retest reliability. The result also indicated that 2007 performance on this measure accounted for 16% of the variance in 2008 scores.

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?*)

Interpretation of measure score reliability testing: The testing suggests the measure has good reliability with beta binomial result of 0.718 exceeding the 0.7 threshold

---

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

2018 Submission

We assessed construct and face validity for this measure.

Method of testing construct validity: We tested for construct validity by exploring whether the Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia measure is correlated with the Diabetes Monitoring for People With Diabetes and Schizophrenia measure. We hypothesized that organizations that perform well on the Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia measure should perform well on the Diabetes Monitoring for People With Diabetes and Schizophrenia because the two measures both focus on patients with schizophrenia and whether their chronic condition, i.e., cardiovascular disease or diabetes, is being monitored.

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

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

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

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

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

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be

effectively collected, reported, and audited before it is used for public accountability or accreditation. This is not testing – the measure was already tested as part of its development – rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

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

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.
Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year's HEDIS Volume 2.

2012 Submission
Validity was assessed using several complementary methods.

Face validity was assessed through a multistakeholder Technical Advisory Group responsible for overseeing measure development. Additionally, face validity was captured through a public comment period and a series of focus groups involving the Medicaid Medical Directors Learning Network, Managed Behavioral Health Care Organizations, and State Mental Health Commissioners and Medical Directors. The panelists assessed the usability and feasibility of the measures.

Concurrent validity was assessed via Medicaid resource utilization from the Medicaid claims data. We examined rates of schizophrenia-related hospital and emergency room utilization as well as total Medicaid costs comparing beneficiaries in the highest and lowest performance quartiles for each measure.

Convergent and discriminant validity were assessed using the Medicaid Analytic Extract (MAX) from Medicaid claims in using 2007 data. Pearson correlation coefficients were used to assess measure correlations. We hypothesized similar measures (e.g. screening and monitoring) would be correlated and (b) process measures would have negative correlations with measures of adverse events (e.g. mental health emergency room utilization).

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

Statistical results of construct validity testing: The results in Table 1 indicate that there is a strong, positive relationship between the Cardiovascular Monitoring for People with Cardiovascular Disease measure and the Diabetes Monitoring for People with Diabetes and Schizophrenia measure. The relationships are statistically significant ($p<0.05$).

Table 1. Correlations in Medicaid Measures – 2016

	Pearson Correlation Coefficient
	Diabetes monitoring for people with diabetes and schizophrenia
Cardiovascular monitoring for people with cardiovascular disease and schizophrenia	0.66

Note: $p<0.05$

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

2012 Submission:
Face validity:
The measures were deemed important, usable, and feasible to collect by the Technical Advisory Group overseeing the measure development, as well as focus groups with the Medicaid Medical Directors Learning Network, Managed Behavioral Healthcare Organizations, and State Mental Health Commissioners and Medical Directors.

Among 22 states, the measure had a minimum value of 11.7%, mean=54.5%, 25th percentile=44.4%, median=59.6%, 75th percentile=67.3% and a maximum value of 85.7%.

Concurrent validity:
Beneficiaries in the lowest performing states for the measure had higher rates of schizophrenia related hospitalization and ED use (24.2% and 26.6%, respectively) than individuals in the highest performing states (17.1% and 16.1%, respectively).

Concurrent and discriminant validity:

Performance on the measure was significantly correlated with the diabetes screening and monitoring measure (r=0.20 and 0.89, respectively).

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
2018 Submission
Interpretation of construct validity testing: The two measures had strong positive correlation, which indicates the measure has good construct validity.

Interpretation of systematic assessment of face validity: NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Cardiovascular monitoring for people with cardiovascular disease and schizophrenia (SMC)* is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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*)
Testing was not performed for exclusions.

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

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

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

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

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)*

2018 Submission

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

2012 Submission

Pearson correlations, means and percentiles are reported.

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

2018 Submission

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicaid	72	77.5	9.9	63.2	72.7	77.6	84.6	88.3	11.9	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

2012 Submission
Among 22 states, the measure had a minimum value of 11.7%, mean=54.5%, 25th percentile=44.4%, median=59.6%, 75th percentile=67.3% and a maximum value of 85.7%.

2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
2018 Submission
The difference between the 25th and 75th percentile is statistically significant for the Medicaid product line. For Medicaid plans, there is a 11.9 percentage point gap between 25th and 75th percentile plans. This gap represents an average 12 more patients with schizophrenia and cardiovascular disease having an LDL-C test during the measurement year in high performing Medicaid plans compared to low performing plans (estimated from average health plan eligible population).

2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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)
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.
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures

3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Specific Plan for Use	Current Use (for current use provide URL)
Not in use	Public Reporting Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Health Plan Ratings https://reportcards.ncqa.org/#/health-plans/list Payment Program Physician Value-Based Payment Modifier (VBM) https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html Quality Improvement (external benchmarking to organizations) Physician Feedback/Quality and Resource Use Reports (QRUR) https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/downloads/QRUR_Presentation.pdf Annual State of Health Care Quality

	http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass

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

PHYSICIAN VALUE-BASED PAYMENT MODIFIER (VBM): This measure is used in the Physician Value-Based Modifier which provides differential payment to a physician or group of physicians under the Medicare Physician Fee Schedule (PFSS). VBM is based on the quality of care provided in comparison to the cost of care within a performance period. The Value Modifier is an adjustment made to Medicare payments for items and services under the Medicare PFS.

NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2017, the report included results from calendar year 2016 for health plans covering over 171 million people.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states were included in the rankings.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

PHYSICIAN FEEDBACK/QUALITY AND RESOURCE USE REPORTS (QRUR): This measure is used in the Physician Feedback Program and Quality and Resource Use Reports which provide comparative performance information to
Medicare Fee-For-Service physicians. The Quality and Resource Use Reports show physicians the portion of their Medicare fee-for-service (FFS) patients who have received indicated clinical services, how patients utilized services, and how Medicare spending for their patients compares to average Medicare spending.

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System.

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.
NCQA measures are evaluated regularly. During this "reevaluation" process, we seek broad input on the measure, including input on performance and implementation experience. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have not reported significant barriers to implementing this measure, as it uses the administrative data collection method. Questions have generally centered around minor clarification of the specifications, including how to identify the eligible population. NCQA responded to all questions to ensure consistent implementation of the specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities.

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.**
Feedback has not required modification to this measure.

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.
From 2015 to 2017, performance rates for this measure have been generally stable. In 2017, Medicaid plans had an average performance rate of 78 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement. In 2017, Medicaid plans in the 10th percentile had a rate of 63 percent, compared to 88 percent among plans in the 90th percentile.

This measure was first introduced in HEDIS 2013. Rates for Medicaid were 67.8 percent. In the last 6 years, we have seen improvement of two percent.

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 were no identified unintended consequences for this measure during testing or since implementation.

4b2.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unintended consequences for this measure during testing or since implementation.

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)**
1932 : Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
1934 : Diabetes Monitoring for People With Diabetes and Schizophrenia (SMD)

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

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.**
N/A

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.)
N/A

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):** National Committee for Quality Assurance
Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-

Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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 Technical Advisory Group advised Mathematica Policy Research, Inc. and the National Committee for Quality Assurance during measure development. The TAG was responsible for providing feedback on measure concepts, specifications, results from field and data testing. The TAG consisted of a multistakeholder group of experts with knowledge in behavioral health and quality measurement.

Technical Advisory Group Roster:
Alisa Busch, MD, MS
Enola Proctor, PhD, MSW
David Shern, PhD
Wilma Townsend, MSW
Dan Ford, MD, MPH
Lorrie Rickman-Jones, PhD
Eric Hamilton
Alexander Young, MD, MHS
Peter Delany, PhD
Ben Druss, MD, MPH
Maureen Corcoran, MSN, MBA
Mike Fitzpatrick, MSW
Anita Yuskauskas
Bob Heinssen, PhD

Consultants:
Lisa Dixon, MD, MPH
Julie Kreyenbul, PharmD, PhD

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers
Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

Behavioral Health Measurement Advisory Panel:
Katharine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute
Christopher Dennis, MD, MBA, FAPA, Landmark Health, LLC
Ben Druss, MD, MPH, Emory University

Frank Ghinassi, PhD, ABPP, Rutgers University Behavioral Health Care
Connie Horgan, ScD, Brandeis University
Laura Jacobus-Kantor, PhD, SAMHSA
Jeffrey Meyerhoff, MD, Optum
Harold Pincus, MD, College of Physicians and Surgeons, Columbia University, New York Presbyterian Hospital, RAND
Michael Schoenbaum, PhD, National Institute of Mental Health
John Straus, MD, Massachusetts Behavioral Health Partnership-A Beacon Health Options Company

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2012

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Every 3-5 years

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

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 "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA. © 2017 by the National Committee for Quality Assurance

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 1934
Measure Title: Diabetes Monitoring for People With Diabetes and Schizophrenia (SMD)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of patients 18 – 64 years of age with schizophrenia and diabetes who had both an LDL-C test and an HbA1c test during the measurement year.
Developer Rationale: The evidence suggests a higher prevalence of diabetes and non-treatment rates for individuals with schizophrenia. Monitoring may lead to proper management for diabetes in this population and may reduce morbidity and mortality

Numerator Statement: One or more HbA1c tests and one or more LDL-C tests performed during the measurement year.
Denominator Statement: Patients age 18-64 years of age as of the end of the measurement year (e.g. December 31) with a schizophrenia and diabetes diagnosis. Patients age 18-64 years of age as of the end of the measurement year (e.g. December 31) with a schizophrenia and diabetes diagnosis.
Denominator Exclusions: Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients who do not have a diagnosis of diabetes (Diabetes Value Set), in any setting, during the measurement year or year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes (Diabetes Exclusions Value Set), in any setting, during the measurement year or the year prior to the measurement year.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan, Integrated Delivery System, Population : Regional and State

Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

Maintenance of Endorsement - Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

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 which shows that patients with schizophrenia and diabetes have a higher prevalence of non-treatment. Therefore, proper monitoring and management of diabetes for those with schizophrenia may reduce mortality and morbidity.
- The developer provides Clinical Practice Guideline recommendation and a systematic review of the evidence including:
    - American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition. **Recommendations within these guidelines range from Grade I (substantial clinical confidence) to Grade II (moderate clinical confidence).**
    - American Diabetes Association (2018). Standards of medical care in diabetes--2018. **Grade E (Expert consensus or clinical experience).**
    - De Hert, M., Vancampfort, D., Correll, C.U., et al. Guidelines for screening and monitoring of cardiometabolic risk in schizophrenia: systematic evaluation (2011). **Grade: Four of the 18 evaluated guidelines are of good quality and should guide clinicians' screening and monitoring practices.**
    - Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. (2016). **This systematic review was conducted in accordance with the Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard.**

- In addition, the developer cites the APA 2009 Guideline Watch which cites additional RCTs and studies that have been completed since the 2004 APA Practice Guidelines for the Treatment of Patients with Schizophrenia that furthers the known link between metabolic side effects and antipsychotics used to treat schizophrenia.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates: The developer provided additional systematic reviews of evidence listed above.

Questions for the Committee:

- *The evidence provided by the developer is updated and directionally the same compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *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 summarized the performance data at the health plan level using HEDIS health plan performance rates from 2015-2017. The data is stratified by year and insurance type.

Measurement Year	# of Plans	Median Denom. Size per plan	Mean	St Dev	10th	25th	50th	75th	90th	Interquartile range
2015	110	132	69.4%	0.1	57.9%	65.7%	69.9%	75.5%	79.3%	9.8
2016	131	135	68.2%	0.1	57.7%	62.7%	68.9%	74.5%	78.2%	11.8
2017	151	159	69.7%	0.1	59.6%	64.4%	70.1%	75.3%	78.8%	10.9

- In the previous review of this measure (2012) the developer provided field tests results to show a performance gap. Among 22 states, the measure had a minimum value of 9.1%, mean=57.3%, 25th percentile=55.6%, median=62.1%, 75th percentile=67.7% and a maximum value of 81.6%.

Disparities

- The developer does not provide disparities data since HEDIS data is stratified by type of insurance. While not specified in this measure, this measure can also be stratified by demographic variables in order to assess the health care disparities.
- The developer provides a summary of research studies demonstrating that individuals with serious mental illness have an increased risk for diabetes as well as disparities in their care.

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:

**The developer provides a logic model which shows that patients with schizophrenia and diabetes have a higher prevalence of non-treatment. Therefore, proper monitoring and management of diabetes for those with schizophrenia may reduce mortality and morbidity. The developer cites well-established practice guidelines and a 2016 meta-analysis.
**Measure applies directly to patient care and reducing poor outcomes from diabetes often associated with SMI. It relies on patients with schizophrenia having one or more HbA1c tests and one or more LDL-C tests performed during the measurement year. There are a higher prevalence of diabetes in this population and therefore measure makes sense. Measure applies directly as it offers early intervention opportunity.
**Evidence to support SIGNIFICANCE of the measure is strong.
**Developer provides evidence of the need for this measure through its logic model, Clinical Guideline Recommendations from 4 separate organizations, and a number of RCTs.
**This measure is critically important for addressing he high prevalence of co-morbidity between schizophrenia and diabetes. Further, individuals with schizophrenia are much more likely be unable to adequately manage their diabetes symptoms because of sedentary lifestyle and the side effects associated with a number of antipsychotic medications. The structure and process of this measure is directly related to the goal of improving the overall health of people with schizophrenia.
**The evidence is good. I have just 2 concerns: which are more addressed below
1. Given some health plans have a low of 9% -I wonder how they have that measure so low, not capturing the data, not able to capture etc.

2. using pharmacy data to get a denominator for diabetics. I noticed metformin alone is not used. As used for weight loss, prevention, prediabetes, etc off- label. I suspect some of the other medications may be similarly headed.

1b. Performance Gap

Comments:

**Yes, 10th percentile performance is at 59.6% and 90th percentile performance is at 78.8%.
**People with schizophrenia have higher prevalence of diabetes that is untreated. Mean testing for diabetes in this population is 68-69% so there are opportunities for improvement.
**Very little change on average/3 years: 68.2%-69.7%. Gap is at best supported by differences in the 10th and 90th percentile for health plans but these appear stable at about 59.6% and 78.8%.
**The evidence provided by the developer shows a rather significant performance gap between states, suggesting that it could be improved with a performance measure.
**Yes, the performance data clearly demonstrates an enormous gap in diabetes care for individuals with schizophrenia. The submission does not include data on subpopulations. We know that there are significant disparities in care for people with and without schizophrenia in diabetes care.
**Yes, I would love to look at more of the demographic of the data between best and worst.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing
2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability; Missing Data
2c. For composite measures: empirical analysis support composite approach

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.

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)?*
- *The 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 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**

Committee pre-evaluation comments

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

2a1. Reliability – Specifications

Comments:

**Data specifications are clear.
**Codes are clearly defined. This could easily be implemented.
**Reliability was tested using the 2016 HEDIS data that included 151 Medicaid plans.
**There is little evidence that this measure cannot be consistently implemented.
**Given some of the lows of 9% I question that data.

2a2. Reliability – Testing

Comments:

**No. Reliability was calculated using the beta binomial method with a result of 0.855 suggesting strong reliability.
**No concerns.
**Based solely on estimates using beta-binomial model. .855 (Medicaid). The main limitation is use of health plan level data given HEDIS measure.
**The Beta-Binomial Statistic was .855, suggesting the measure has very good reliability.
**No. The measure is repeatable and should produce the same results when assessed in the same population in the same time period. It is also precise enough to distinguish differences in performance across providers.

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

Comments:

**Both construct validity and face validity were assessed. Construct validity had a pearson correlation of 0.66 with the corresponding measure of cardiovascular monitoring of individuals with schizophrenia and bipolar disorder and cardiovascular disease. To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans' performance is significantly different from each other. Variation was demonstrated with the 10th percentile performance at 59.6% compared to the 90th percentile performance at 78.8%.
**Construct validity based only on examining correlation in adherence to a conceptually similar HEDIS measure (Schiz/BPD using AP Meds). Pearson correlation=.66. Face validity results not presented, but stated the measure has face validity given "input" from advisory panel and public comments. NCQA measure development process described.
**No concerns with validity and should produce comparable results. Missing data in no way threatens validity.

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

Comments:

**No capacity to adjust for social risk factors. Little capacity to stratify by demographic characteristics unless "data available to the plan"
**The measure incorporates only individuals with schizophrenia already diagnosed with diabetes. The existence of the measure will hopefully spur psychiatrists and mental health professionals to more carefully address risk factors associated with diabetes -- weight, blood pressure, etc.
**Including folks that don't truly have diabetes given using pharmacy data. Also, we are learning more about "curing" diabetes through loss (often through surgery) or the rare individual that does it on their own and the "chart lore" effect. Also important to our folks is the atypical medication DM.

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 defined fields in electronic claims.
- No fees or licensure requirements are required.
- The developer notes that the measure has clear specifications but data methods and calculation methods may vary. Therefore, NCQA conducts an independent audit of all HEDIS collection and reporting processes as well as an audit of the data which are manipulated by those processes in order to verify that HEDIS specifications are met.

Questions for the Committee:

- *Does the Committee have any concerns in regards to the feasibility of this measure based on endorsement maintenance updates?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**All data elements are in defined fields in electronic claims. Additionally, NCQA has a HEDIS audit process and a system of providing real time feedback to users.
**Data collection is easily pulled from claims/encounter data and outcomes are easily defined.
**Highly feasible using electronic data sources.
**All data elements are available in electronic fields in claims data. This measure could feasibly be implemented without much burden to providers. No fees are associated with this measure.
**According to the submission, all of the data elements are electronic form. This is a fairly simple process measure, either diabetes is being monitored or it isn't. Changing behavior -- diet, exercise, adherence to diabetes treatment -- is a much larger public health challenge.

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**

Accountability program details:

- Physician Value-Based Payment Modifier
- NCQA State of Health Care Quality Report

- NCQA Health Plan Ratings/Report Card
- NCQA Quality Compass
- Physician Feedback/Quality and Resource Use Reports

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

- The developer publicly reports rates across all plans and creates benchmarks to help plans how they perform compared to other plans.
- The developer publishes performance results and data annually in their Quality Compass tool and presents data at various conferences and webinars. The developer also provides regular technical assistance through its Policy Clarification Support System.
- The developer uses several methods to obtain input from users during its "reevaluation process," including, vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer noted that the health plans have not reported significant implementation barriers. Questions from users typically center around clarifications of the specifications such as confirmation that patients are correctly excluded from the measure.

Additional Feedback:

- 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

- The developer notes that in the past 2 years, performance rates for this measure have been generally stable or shown slight improvement. In 2017, Medicaid plans had an average performance rate of 70 percent. The most significant variation is between the 10th and 90th percentiles, suggesting room for improvement.
- This measure was first introduced in HEDIS 2013. Rates for Medicaid were 67.8 percent. In the last 6 years, the developer has seen an improvement of 2 percent.

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 were reported by the developer

Potential harms

- None were reported by the developer

Additional Feedback:

- N/A

***Questions for the Committee**:*

o *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:

**NCQA measures are evaluated regularly. During this "reevaluation" process, we seek broad input on the measure, including input on performance and implementation experience. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility. In general, health plans have not reported significant barriers to implementing this measure, as it uses the administrative data collection method. Questions have generally centered around minor clarification of the specifications, such as confirmation that patients are correctly excluded from the measure according to the measure specification. NCQA responded to all questions to ensure consistent implementation of the specifications.

**Measure is currently used by NCQA and providers in public reporting, the annual state of health care quality report card, etc... NCQA publically reports rates across all plants in order to help plans understand how they perform in relation to other plans.

**According to the submission, this measure is being collected and publicly reported across Physician Value-Based Payment Modifier and multiple NCQA reporting systems. All of these systems have significant feedback processes for providers around measure performance and implementation.

4b1. Usability – Improvement

Comments:

**No perceived harm. Measure would allow for opportunities for improved diabetes care reducing need for more serious, costly, and challenging intervention down the line.

**There may be some difficulty interpreting findings if persons with SMI on AP meds did not get screened and persons with schizophrenia and prescribed oral hypoglycemic (compliance?) get screened, do we "know" if the person is being treated for DM due to metabolic syndrome because of poor ap med safety monitoring? I understand that this is not at the health plan level as developed, but I would anticipate that some might use this measure at the patient level.

**There are no reported unintended consequences reported for this measure.

**Collecting data on diabetes management in this population is critical public health priority. It is essential to improving the health of people with schizophrenia and addressing early mortality. Any unintended consequences are far outweighed by the potential public health benefit.

Criterion 5: Related and Competing Measures

Related or competing measures

- 1932: Diabetes Screening for People with Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Mediations (SSD)
- 1933: Cardiovascular Monitoring for People with Cardiovascular Disease and Schizophrenia (SMC)

Harmonization

- Specifications are harmonized to the extent possible, per the developer.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 1934

Measure Title: Diabetes Monitoring for People With Diabetes and Schizophrenia (SMD)

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document

   ***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.

The measure is specified and tested at the health plan level

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

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐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, skip Questions #3-8, then go to Question #9)

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

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

   Reliability of the measure score was assessed using 2016 HEDIS data that included 151 Medicaid plans.

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

   The developer used a beta-binomial model to calculate the plan score. Results of reliability testing was 0.855.

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #6)

   ☒Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

   Beta-Binomial Statistic:

Medicaid
0.855

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

   REFERENCE: Testing attachment, section 2a2.

***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

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

REFERENCE: Testing attachment, section 2a2.2

***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 #8)

☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?

REFERENCE: Testing attachment, section 2a2

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

☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)

☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)

☐Insufficient (go to Question #9)

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

REFERENCE: testing attachment section 2b1.

NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP:** You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but **check with NQF staff before proceeding, to verify.

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

***TIPS**: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☐Yes (go to Question #12)

☒No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

No concerns were identified.

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

REFERENCE: Testing attachment, section 2b2.

***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 #13)

☒No (go to Question #13)

☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

Testing was not performed for exclusions

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

N/A Process measure

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

REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

The developer compared performance between to randomly selected plans at the 25th and 75th percentile to understand the variation in performance. Using the t-test method, they calculated a testing statistic based

on the sample size, performance rate, and standardized error of each plan, which was then compared against a normal distribution. The results showed that the two plans' performance was significantly different from each other.

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicaid	278	69.7	7.9	59.6	64.4	70.1	75.3	78.8	10.9	<0.05

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☐No (go to Question #16)
☒Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☐No (please explain below, then skip Questions #18-23 and go to Question #24)

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☒Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.

***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 #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)

☐Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

To assess the validity of the measure, the developer conducted construct validity testing using the Pearson correlation coefficient to examine the association between using this measure and measure 1932, which both focus on patients with schizophrenia and whether they received care for diabetes. They found that there is a statistically significant (0.66) and positive relationship between the two measures.

21. Was validity testing conducted with <u>patient-level data elements</u>?

REFERENCE: Testing attachment, section 2b1.

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

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

***REFERENCE:** Testing attachment, section 2b1.*

***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 #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has <u>not</u> been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. Was <u>face validity</u> 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***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 #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☒ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1934

Measure Title: Diabetes Monitoring for People With Diabetes and Schizophrenia (SMD)

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

Date of Submission: 4/2/2018

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: Diabetes Monitoring for People With Diabetes and Schizophrenia

☐ 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.

Patients with schizophrenia and diabetes>>higher prevalence of non-treatment rates for individuals with schizophrenia>>diabetes monitoring>>proper management of diabetes>>reduced morbidity and mortality

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

****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

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. APA Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition; 2004 Feb. 184 p. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf and GUIDELINE WATCH: PRACTICE GUIDELINE FOR THE TREATMENT OF PATIENTS WITH SCHIZOPHRENIA; American Psychiatric Association, 2009 SEP. 10 P. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia-watch.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.	**Acute Phase Treatment [A, A-, B, C, D, E, F, G]** • General medical health as well as medical conditions that could contribute to symptom exacerbation can be evaluated by medical history, physical and neurological examination, and appropriate laboratory, electrophysiological, and radiological assessments [I]. Measurement of body weight and vital signs (heart rate, blood pressure, temperature) is also recommended [II]. • Other laboratory tests to be considered to evaluate health status include a complete blood count (CBC); measurements of blood electrolytes, glucose, cholesterol, and

	triglycerides; tests of liver, renal, and thyroid function; a syphilis test; and when indicated and permissible, determination of HIV status and a test for hepatitis C [II]. **Stable Phase [A, A-, B, C, D, E, F, G]** • Routine monitoring for obesity-related health problems (e.g., high blood pressure, lipid abnormalities, and clinical symptoms of diabetes) and consideration of appropriate interventions are recommended particularly for patients with BMI in the overweight and obese ranges [II]. Clinicians may consider regular monitoring of fasting glucose or hemoglobin A1c levels to detect emerging diabetes, since patients often have multiple risk factors for diabetes, especially patients with obesity [I]
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline. Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are “blind” to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.

	[C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	[I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence.
Provide all other grades and definitions from the recommendation grading system	[III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	"Relevant literature was identified through a computerized search of PubMed for the period from 1994 to 2002. Using the keywords schizophrenia OR schizoaffective, a total of 20,009 citations were found. After limiting these references to clinical trials and meta-analyses published in English that included abstracts, 1,272 articles were screened by using title and abstract information. The Cochrane Database of Systematic Reviews was also searched by using the keyword schizophrenia. Additional, less formal literature searches were conducted by APA staff and individual members of the work group on schizophrenia. Sources of funding were considered when the work group reviewed the literature but are not identified in this document. When reading source articles referenced in this guideline, readers are advised to consider the sources of funding for the studies"

Estimates of benefit and consistency across studies	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline"
What harms were identified?	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	To our knowledge, there have been no published studies since the clinical practice guidelines that would contradict the current body of evidence.

Table 2. ADA Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**2018 Submission** American Diabetes Association (2018). Standards of medical care in diabetes--2018. Diabetes Care, 41, S28–S37. http://care.diabetesjournals.org/content/diacare/suppl/2017/12/08/41.Supplement_1.DC1/DC_41_S1_Combined.pdf **2012 Submission** American Diabetes Association (2011). Standards of medical care in diabetes--2011. Diabetes Care, 34, S11-61. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3006050/
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.	**2018 Submission** • Perform the A1C test at least two times a year in patients who are meeting treatment goals (and who have stable glycemic control). E • Obtain a lipid profile at initiation of statins or other lipid-lowering therapy, 4–12 weeks after initiation or a change in dose, and

	annually thereafter as it may help to monitor the response to therapy and inform adherence. E **2012 Submission** Perform the A1c test at least two times a year in patients who are meeting treatment goals (and who have stable glycemic control). Grade E Recommendation.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2018 Submission** **E:** Expert consensus or clinical experience **2012 Submission** E: Expert consensus or clinical experience
Provide all other grades and definitions from the evidence grading system	**2018 Submission** **A:** Clear evidence from well-conducted, generalizable randomized controlled trials that are adequately powered, including • Evidence from a well-conducted multicenter trial • Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., "all or none" rule developed by the Centre for Evidence-Based Medicine at the University of Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including • Evidence from a well-conducted trial at one or more institutions • Evidence from a meta-analysis that incorporated quality ratings in the analysis B: Supportive evidence from well-conducted cohort studies • Evidence from a well-conducted prospective cohort study or registry • Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study **C:** Supportive evidence from poorly controlled or uncontrolled studies

	• Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results • Evidence from observational studies with high potential for bias (such as case series with comparison with historical controls) • Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation **2012 Submission** **A:** Clear evidence from well-conducted, generalizable, randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted multicenter trial • Evidence from a meta-analysis that incorporated quality ratings in the analysis Compelling nonexperimental evidence, i.e., "all or none" rule developed by Center for Evidence Based Medicine at Oxford Supportive evidence from well-conducted randomized controlled trials that are adequately powered, including: • Evidence from a well-conducted trial at one or more institutions • Evidence from a meta-analysis that incorporated quality ratings in the analysis **B:** Supportive evidence from well-conducted cohort studies • Evidence from a well-conducted prospective cohort study or registry • Evidence from a well-conducted meta-analysis of cohort studies Supportive evidence from a well-conducted case-control study **C:** Supportive evidence from poorly controlled or uncontrolled studies • Evidence from randomized clinical trials with one or more major or three or more minor methodological flaws that could invalidate the results • Evidence from observational studies with high potential for bias (such as case series with comparison to historical controls)

	• Evidence from case series or case reports Conflicting evidence with the weight of evidence supporting the recommendation
Grade assigned to the **recommendation** with definition of the grade	**2018 Submission** No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. **2012 Submission** N/A
Provide all other grades and definitions from the recommendation grading system	**2018 Submission** No additional grading was provided, grades assigned to evidence is the same with grades assigned to recommendations. **2012 Submission** N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2018 Submission** The ADA does not provide information on the systematic review conducted to support its guideline and the recommendations mentioned above. In lieu of the ADA systematic review, we provide information on an additional systematic review that supports the ADA's recommendations in Table 2. **2012 Submission** 7; This measure is supported by prevalence studies that show a higher prevalence rate of diabetes for individuals with schizophrenia
Estimates of benefit and consistency across studies	**2018 Submission** See Tables 3 and 4. **2012 Submission** • Benefit: Monitoring allows for the ability to treat appropriately, if warranted. Given the long asymptomatic period in the early natural history of diabetes, within patients with schizophrenia, proportion of people with undiagnosed diabetes is much higher. • Cost: LDL-C and HbA1c test

	• The studies consistently show that individuals with schizophrenia have a higher prevalence of diabetes.
What harms were identified?	**2018 Submission** See Tables 3 and 4. **2012 Submission** N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2018 Submission** To our knowledge, there have been no published studies since the clinical practice guidelines that would impact the recommendations. **2012 Submission** N/A

Table 3. Systematic Review supporting Diabetes Monitoring for People With Diabetes and Schizophrenia

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	De Hert, M., Vancampfort, D., Correll, C.U., Mercken, V., Peuskens, J., Sweers, K., van Winkel, R., Mitchel, A.J. 2011. Guidelines for screening and monitoring of cardiometabolic risk in schizophrenia: systematic evaluation. The British Journal of Psychiatry (2011) 199, 99–105. https://pdfs.semanticscholar.org/bef2/d8f81c9c99c5057906a5f75fd9bb03a93b15.pdf
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	"The aim of this study was to perform a systematic review of the available clinical practice guidelines for the screening and monitoring of cardiometabolic risk in people with schizophrenia and related psychotic disorders. The quality of these guidelines is assessed with the Appraisal of Guidelines for Research and Evaluation (AGREE)." Based on this review of the guidelines, a monitoring protocol for managing cardiovascular disease risk in patients in clinical practice is proposed. Those who already present with cardiovascular risk factors should be monitored frequently. At the start of a new treatment, assessments should be repeated 6 and 12

	weeks after initiation of the new antipsychotic drug treatment.
Grade assigned for the quality of the quoted evidence with definition of the grade	Four of the evaluated guidelines are of good quality and should guide clinicians' screening and monitoring practices
Provide all other grades and definitions of the evidence in the grading system	The evaluation and comparison of the guidelines consisted of "23 items grouped in six domains: scope and purpose; rigour of development; stakeholder involvement; clarity and presentation; applicability; and editorial independence. Each item is scored on a 4-point scale (strongly agree, agree, disagree and strongly disagree) with proposed anchor points to evaluate in which way the guideline fulfils the domain. The scores are standardised in a percentage score that enables comparison between guidelines (obtained score-minimum possible score)/ (maximum possible score-minimum possible score). The final component of the AGREE instrument involves a recommendation regarding the use of the guidelines in practice as 'recommended', 'recommended (with provisos or exceptions)', 'would not recommend or unsure', depending on the number of items and domains if the score was 460%, 30–60% and 530%, respectively. Three raters (D.V., K.S. and M.D.H.) independently scored the identified guidelines (M.D.H. acknowledges a potential conflict of interest because he co-authored two of the assessed guidelines). A mean score was calculated for each item from which the percentage score was derived according to the AGREE manual. In addition, each guideline was independently evaluated regarding the specific content and scope of what should be monitored by whom. Process indicators were predefined and scored on a standardised scoring sheet (online supplement 1). Intraclass correlation coefficients (ICC) with a 95% confidence interval were calculated as an overall indicator of agreement among the raters for each of the 23 items of the AGREE instrument."
What is the time period covered by the body of evidence?	1 January 2000 until 1 April 2010
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity: 18 Quality: "A total of 18 unique guidelines were identified for AGREE evaluation either from the USA (2), Australia (2), Brazil (1), Canada (1) or Europe (12), and all were published between 2004

	and 2010. All papers covered diabetes and cardiovascular disease risk in individuals treated with antipsychotic agents, whereas some had a broader scope also including other physical health domains and other side-effects. Regarding the domain rigour of development all except one guideline had a score below 50%. Although some guidelines presented data from a systematic review of the literature, the search strategy for literature selection was missing in all but one guideline. Only two guidelines presented levels/quality of the evidence and one presented meta-analytic data. More than half (61%) of the guidelines were developed with a consensus model. Within this domain the criterion about the updating of the recommendation was not fulfilled by any of the guidelines. The older UK guideline has a low score on this item, but the paper was published in a themed issue of the journal, with different papers presenting a systematic review of the literature in that same issue. Scores in the application domain were satisfactory in five guidelines. The guidelines with a low score on this domain failed to discuss the organisational aspects of introducing screening and monitoring. Health economic aspects were mentioned in some guidelines but the additional cost of screening and monitoring was explicitly available in only one."
What is the overall quality of evidence across studies in the body of evidence?	Overall, the quality of evidence supporting this measure is strong. There are 18 guidelines in the evidence review that examine the effectiveness of cardiovascular screening monitoring for individuals with schizophrenia.
Estimates of benefit and consistency across studies in body of evidence– what are the estimates of benefits?	"Clinical practice guidelines are considered a good option for translating research into clinical practice. They are defined as 'systematically developed statements to assist practitioner and patient decisions about appropriate healthcare for specific clinical circumstances'. Their potential to improve patient care and outcomes depends largely on the quality and independence of the guideline. Recommendations may be biased because of non-systematic selection, inadequate interpretation or lack of scientific evidence. The content may initially be decided through consensus, whereas scientific evidence to support the consensus is added afterwards. The influence of the context within which the guidelines are produced (for example by medical societies or with support of

	pharmaceutical companies) has also been mentioned in relation to the variation across guidelines. Quality evaluations have recently been performed for other diseases in relation to metabolic and cardiovascular risk monitoring. Similar to our findings, for diabetes and cardiovascular disease the rigour of development and other quality indications, such as stakeholder involvement and editorial independence, were not ideal in a number of these guidelines. This was the case, despite medical societies developing stringent methodologies for these diseases according to internal guidelines/procedures. Moreover, editorial independence was also often a problematic area, and frequently guidelines were not based on high-quality evidence."
What harms were studied and how to they affect the net benefit (benefits over harm)?	No harms associated with screening and monitoring were identified in the evidence reviewed.

Table 4. Systematic Review supporting Diabetes Monitoring for People With Diabetes and Schizophrenia

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. World Psychiatry. 2016;15(2):166-174. https://doi.org/10.1002/wps.20309
What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?	"This meta-analysis aimed: a) to describe pooled frequencies of T2DM in people with SMI; b) to analyze the influence of demographic, illness and treatment variables as well as T2DM assessment methods (blood testing, self-report, charts); and c) to describe T2DM prevalence in studies directly comparing persons with each specific SMI diagnosis to general population samples… In patients with T2DM (and those with pre-diabetes), fasting blood glucose and HBA1c should be measured more frequently (approximately every 3-6 months). An annual examination should include measurement of CVD risk factors, glomerular filtration rate and albumin to creatinine ratio, an eye examination, ideally including fundus photography, and foot examination to diagnose early signs of complications"

Grade assigned for the quality of the quoted evidence with definition of the grade	This systematic review was conducted in accordance with the M eta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614) and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097)
Provide all other grades and definitions of the evidence in the grading system	This systematic review was conducted in accordance with the M eta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614) and in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standard (http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097)
What is the time period covered by the body of evidence?	Database inception to August 1, 2015
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: 118 Quality of studies: “observational studies (cross-sectional, retrospective and prospective studies) and randomized controlled trials in adults with a psychiatric diagnosis of schizophrenia or related psychotic disorders, bipolar disorder or MDD according to the DSM-IV-TR or the ICD-10, irrespective of clinical setting (inpatient, outpatient or mixed, community setting), that reported study-defined T2DM prevalences.”
What is the overall quality of evidence across studies in the body of evidence?	Overall, the quality of evidence supporting this measure is strong. There are over 100 studies in the evidence review that examine the prevalence and effectiveness of diabetes screening and monitoring for individuals with SMI, including schizophrenia. Further, the quality of studies included in the systematic review were well-designed randomized control trials and observational studies and.

Estimates of benefit and consistency across studies in body of evidence– what are the estimates of benefits?	"To our knowledge, this is the first meta-analysis of T2DM including and comparing data from the three main SMIs, namely schizophrenia and related psychotic disorders, bipolar disorder and MDD. Approximately one in 10 individuals with SMI (11.3%; 95% CI: 10.0%-12.6%) had T2DM, and the relative risk for T2DM in multi-episode persons with SMI was almost double (RR=1.85, 95% CI: 1.45-2.37) that found in matched general population comparison samples. T2DM prevalences were consistently elevated for each of the three diagnostic subgroups compared to the general population, and comparative meta-analyses found no significant differences across schizophrenia, schizophrenia spectrum disorders, bipolar disorder and MDD. Thus, other diagnostic-independent factors likely influence T2DM frequency, including hyperglycaemia following psychotropic medication use and long-term exposure to unhealthy lifestyle behaviors, as well as potential genetic factors linking psychiatric and medical risk. Knowledge of factors associated with a high T2DM risk can help identify individuals at greatest need for intensive monitoring and intervention. In contrast with general population studies, we found that women with SMI had a higher risk for developing T2DM than men. This finding warrants further investigation, but may be related to a greater propensity to obesity and central obesity in women with SMI compared to men, since central obesity is a significant risk factor for hyperglycaemia. On the other hand, only a minority of analyzed studies did provide information about the mean age in women and men, and it is possible that women with schizophrenia were older, which could have confounded the results. There were no significant differences between the various treatment settings, and data collection before versus after the year 2000. There was also no difference in T2DM prevalence between population based and non-population based studies. In contrast, a higher T2DM prevalence was observed in studies relying upon clinical data gleaned from file and chart reviews versus self-report studies. A trend for higher T2DM was found in retrospective studies

	versus cross-sectional (p=0.054) and versus prospective (p=0.053) studies.”
What harms were studied and how to they affect the net benefit (benefits over harm)?	No harms associated with testing were identified in the evidence reviewed.

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 APA 2009 Guideline Watch identified a number of controlled clinical trials examining treatments to prevent or treat weight gain and metabolic changes caused by antipsychotic use. The Guideline Watch additionally cite several randomized control trials (RCTs) related to new antipsychotics used to treat schizophrenia. This report highlights research studies published since the 2004 APA Practice Guidelines for the Treatment of Patients with Schizophrenia and furthers the known link between metabolic side effects and antipsychotics used to treat schizophrenia.

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

“This watch highlights key research studies published since that date. The studies were identified by a MEDLINE literature search for meta-analyses and randomized, controlled trials published between 2002 and 2008, using the same key words used for the literature search performed for the 2004 guideline.”

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

GUIDELINE WATCH: PRACTICE GUIDELINE FOR THE TREATMENT OF PATIENTS WITH SCHIZOPHRENIA; American Psychiatric Association, 2009 SEP. 10 P.

https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia-watch.pdf

NATIONAL QUALITY FORUM

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 1934
Corresponding Measures:
De.2. **Measure Title:** Diabetes Monitoring for People With Diabetes and Schizophrenia (SMD)
Co.1.1. **Measure Steward:** National Committee for Quality Assurance
De.3. **Brief Description of Measure:** The percentage of patients 18 – 64 years of age with schizophrenia and diabetes who had both an LDL-C test and an HbA1c test during the measurement year.
1b.1. **Developer Rationale:** The evidence suggests a higher prevalence of diabetes and non-treatment rates for individuals with schizophrenia. Monitoring may lead to proper management for diabetes in this population and may reduce morbidity and mortality

S.4. **Numerator Statement:** One or more HbA1c tests and one or more LDL-C tests performed during the measurement year.
S.6. **Denominator Statement:** Patients age 18-64 years of age as of the end of the measurement year (e.g. December 31) with a schizophrenia and diabetes diagnosis.Patients age 18-64 years of age as of the end of the measurement year (e.g. December 31) with a schizophrenia and diabetes diagnosis.
S.8. **Denominator Exclusions:** Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients who do not have a diagnosis of diabetes (Diabetes Value Set), in any setting, during the measurement year or year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes (Diabetes Exclusions Value Set), in any setting, during the measurement year or the year prior to the measurement year.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Health Plan, Integrated Delivery System, Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Nov 02, 2012 **Most Recent Endorsement Date:** Nov 02, 2012

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable.

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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
1934_SMD_MEF_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.
Yes

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 evidence suggests a higher prevalence of diabetes and non-treatment rates for individuals with schizophrenia. Monitoring may lead to proper management for diabetes in this population and may reduce morbidity and mortality

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 following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data are summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data are stratified by year and product line (i.e. Medicaid).

Diabetes Monitoring for People With Diabetes and Schizophrenia– Medicaid Rate (HMO and PPO Combined)
MEASUREMENT YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range
2015 | 69.4% | 0.1 | 57.9% | 65.7% | 69.9% | 75.5% | 79.3% | 9.8
2016 | 68.2% | 0.1 | 57.7% | 62.7% | 68.9% | 74.5% | 78.2% | 11.8
2017 | 69.7% | 0.1 | 59.6% | 64.4% | 70.1% | 75.3% | 78.8% | 10.9

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 47 million Medicaid health plan beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the mean eligible population for the measure across health plans.

Diabetes Monitoring for People With Diabetes and Schizophrenia– Medicaid (HMO and PPO Combined)
YEAR | N Plans | Median Denominator Size per plan
2015 | 110 | 132
2016 | 131 | 135
2017 | 151 | 159

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.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and

follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities. Based on extensive work by NCQA to understand how to promote culturally and linguistically appropriate services among plans and providers, we have many examples of how health plans have used HEDIS measures to design quality improvement programs to decrease disparities in care.

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

A number of research studies, including several meta-analyses, demonstrate that individuals with serious mental illness have an increased risk for diabetes as well as disparities in their care.

One review article estimated the prevalence of diabetes among individuals with SMI is approximately 12% (Holt and Mitchell, 2015), while the prevalence in the general population is approximately 9% aged =18 (CDC, 2017). Additionally, there is a known link between SMI treatments such as anticonvulsants and antipsychotic medications to adverse metabolic risks in patients, such as diabetes (Vancampfort, 2016).

A systematic review article assessed 118 cross-sectional, retrospective and prospective studies, and population versus non-population based studies comparing SMI individuals with non-serious mental illness control groups. Based on this evidence review, authors conclude that diabetes is more common among patients with SMI with a relative risk of 2.04 in patients with schizophrenia or related psychotic disorders compared to the general population (Vancampfort, 2016).

Evidence suggests that individuals with SMI, specifically those with schizophrenia, are at increased risk of developing diabetes due to a higher prevalence of risk factors including tobacco use, poor nutrition and obesity and weight gain from the use of antipsychotics (Mangurian, 2016). Furthermore, these risk factors result in increased morbidity, such as hospitalizations and complications from diabetes, and mortality in the SMI population (Mai et al., 2011; CDC, 2010).

Despite these risks, people with SMI and diabetes receive less ongoing diabetes monitoring and had higher risk for diabetes complications and diabetes-related mortality compared to non-mental health patients (Mai, 2011). In one study analyzing data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia study, researchers found that the rate of non-treatment for individuals with schizophrenia and diabetes was approximately 30% (Nasrallah et al., 2006).

In an additional study examining a national cardiometabolic screening program for 10,084 patients attending mental health clinics, approximately 56% of patients with schizophrenia and metabolic syndrome were not receiving treatment for any metabolic syndrome component (Correll et al., 2010)

References

Banta JE, Morrato EH, Lee SW, et al. (2009) Retrospective Analysis of Diabetes Care in California Medicaid Patients with Mental Illness. J Gen Intern Med. 24:802-8.

Centers for Disease Control and Prevention (CDC). (2010) Diagnosed and undiagnosed diabetes in the United States, all ages, 2010. Retrieved from: http://www.cdc.gov/diabetes/pubs/estimates11.htm. Accessed on June 19, 2014.

Centers for Disease Control and Prevention (CDC). National Diabetes Statistics Report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2017.

Correll, C.U., Druss, B. G., Lombardo, I., O'Gorman, C., Harnett, J.P., Sanders, K.N., Alvir, J.M., Cuffel, B.J. (2010). Findings of a U.S. National Cardiometabolic Screening Program Among 10,084 Psychiatric Outpatients. Psychiatric Services. 2010 Sep;61(9)

Holt R.I., Mitchell A.J. (2015). Diabetes mellitus and severe mental illness: mechanisms and clinical implications. Nat Rev Endocrinol. 2015 Feb;11(2):79-89. doi: 10.1038/nrendo.2014.203.

Mai Q, Holman CD, Sanfilippo FM, et al. (2011) Mental illness related disparities in diabetes prevalence, quality of care and outcomes: a population-based longitudinal study. BMC Medicine. 9:118.

Mangurian, C., Newcomer, J.W., Modlin, C. et al. Diabetes and Cardiovascular Care Among People with Severe Mental Illness: A Literature Review. J GEN INTERN MED (2016) 31: 1083. https://doi.org/10.1007/s11606-016-3712-4

Nasrallah, H.A., Meyer, J.M., Goff, D.C., McEvoy, J.P., Davis, S.M., Stroup, T.S., Lieberman, J.A. Low rates of treatment for hypertension, dyslipidemia and diabetes in schizophrenia: Data from the CATIE schizophrenia trial sample at baseline. Schizophrenia Research 86 (2006) 15-22. https://www.ncbi.nlm.nih.gov/pubmed/16884895

Vancampfort D, Correll CU, Galling B, et al. Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis. World Psychiatry. 2016;15(2):166-174. doi:10.1002/wps.20309.

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):*
Behavioral Health, Endocrine : Diabetes

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

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Populations at Risk : Individuals with multiple chronic conditions

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 Applicable

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:** 1934_SMD_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.
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.
No important changes since the last update.

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).
One or more HbA1c tests and one or more LDL-C tests performed during the measurement year.

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).
An HbA1c test (HbA1c Tests Value Set) and an LDL-C test (LDL-C Tests Value Set) performed during the measurement year (on the same or different dates of service), as identified by claim/encounter or automated laboratory data. The patient must have both tests to be included in the numerator. The organization may use a calculated or direct LDL.

See corresponding Excel document for the LDL-C Tests Value Set and the HbA1c Tests Value Set

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients age 18-64 years of age as of the end of the measurement year (e.g. December 31) with a schizophrenia and diabetes diagnosis.Patients age 18-64 years of age as of the end of the measurement year (e.g. December 31) with a schizophrenia and diabetes diagnosis.

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).
Follow the steps below to identify the eligible population.

Step 1: Identify members with schizophrenia as those who met at least one of the following criteria during the measurement year:

- At least one acute inpatient encounter, with any diagnosis of schizophrenia. Either of the following code combinations meets criteria:
  - BH Stand Alone Acute Inpatient Value Set with Schizophrenia Value Set.
  - BH Acute Inpatient Value Set with BH Acute Inpatient POS Value Set with Schizophrenia Value Set.

- At least two visits in an outpatient, intensive outpatient, partial hospitalization, ED or nonacute inpatient setting, on different dates of service, with any diagnosis of schizophrenia. Any two of the following code combinations meet criteria:
  - BH Stand Alone Outpatient/PH/IOP Value Set with Schizophrenia Value Set.
  - BH Outpatient/PH/IOP Value Set with BH Outpatient/PH/IOP POS Value Set with Schizophrenia Value Set.
  - ED Value Set with Schizophrenia Value Set.
  - BH ED Value Set with ED POS Value Set with Schizophrenia Value Set.
  - BH Stand Alone Nonacute Inpatient Value Set with Schizophrenia Value Set.
  - BH Nonacute Inpatient Value Set with BH Nonacute Inpatient POS Value Set with Schizophrenia Value Set.

Step 2 Identify members from step 1 who also have diabetes. There are two ways to identify members with diabetes: by claim/encounter data and by pharmacy data. The organization must use both methods to identify the eligible population, but a member need only be identified by one to be included in the measure. Members may be identified as having diabetes during the measurement year or the year prior to the measurement year.

Claim/encounter data. Members who met any of the following criteria during the measurement year or the year prior to the measurement year (count services that occur over both years):

- At least two outpatient visits (Outpatient Value Set), observation visits (Observation Value Set), ED visits (ED Value Set) or nonacute inpatient encounters (Nonacute Inpatient Value Set), on different dates of service, with a diagnosis of diabetes (Diabetes Value Set). Visit type need not be the same for the two encounters.
- At least one acute inpatient encounter (Acute Inpatient Value Set), with a diagnosis of diabetes (Diabetes Value Set).

Pharmacy data. Members who were dispensed insulin or oral hypoglycemics/antihyperglycemics on an ambulatory basis during the measurement year or the year prior to the measurement year (Diabetes Medications List).

(See corresponding Excel document for the above value sets)

PRESCRIPTIONS TO IDENTIFY PATIENTS WITH DIABETES (Diabetes Medications List):
Alpha-glucosidase inhibitors:
Acarbose, Miglitol

Amylin analogs:
Pramlinitide

Antidiabetic combinations:
Alogliptin-metformin, Alogliptin-pioglitazone, Canagliflozin-metformin, Dapagliflozin-metformin, Empaglifozin-linagliptin, Empagliflozin-metformin, Glimepiride-pioglitazone, Glimepiride-rosiglitazone, Glipizide-metformin, Glyburide-metformin, Linagliptin-metformin, Metformin-pioglitazone, Metformin-repaglinide, Metformin-rosiglitazone, Metformin-saxagliptin, Metformin-sitagliptin, Sitagliptin-simvastatin
Insulin:
Insulin aspart, Insulin aspart-insulin aspart protamine, Insulin degludec, Insulin detemir, Insulin glargine, Insulin glulisine, Insulin isophane human, Insulin isophane-insulin regular, Insulin lispro, Insulin lispro-insulin lispro protamine, Insulin regular human, Insulin human inhaled

Meglitinides:
Nateglinide, Repaglinide

Glucagon-like peptide-1 (GLP1) agonists:
Dulaglutide, Exenatide, Liraglutide, Albiglutide

Sodium glucose cotransporter 2 (SGLT2) inhibitor:
Canagliflozin, Dapagliflozin, Empagliflozin

Sulfonylureas:
Chlorpropamide, Glimepiride, Glipizide, Glyburide, Tolazamide, Tolbutamide

Thiazolidinediones:
Pioglitazone, Rosiglitazone

Dipeptidyl peptidase-4 (DDP-4) inhibitors:
Alogliptin, Linagliptin, Saxagliptin, Sitaglipin

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.

Exclude patients who do not have a diagnosis of diabetes (Diabetes Value Set), in any setting, during the measurement year or year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes (Diabetes Exclusions Value Set), in any setting, during the measurement year or the year prior to the measurement year.

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.)*

Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began. These patients may be identified using various methods, which may include but are not limited to enrollment data, medical record or claims/encounter data (Hospice Value Set).

Optional exclusion: Exclude patients who do not have a diagnosis of diabetes (Diabetes Value Set), in any setting, during the measurement year or year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes (Diabetes Exclusions Value Set), in any setting, during the measurement year or the year prior to the measurement year.

If a member was identified as a diabetic based on claim or encounter data, as described in step 2 of S.7, the optional exclusions do not apply because the member had a diagnosis of diabetes.

See corresponding Excel document for the value sets referenced above.

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.)*
None.

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. Determine the eligible population: identify patients 18-64 years of age by the end of the measurement year
Step 2. Search for an optional exclusion in the patient's history: Exclude patients from the eligible population if the eligible population if they meet the following criteria:
- Exclude patients who use hospice services or elect to use a hospice benefit any time during the measurement year, regardless of when the services began.
- Exclude patients who do not have a diagnosis of diabetes during the measurement year or year prior to the measurement year and who had a diagnosis of gestational diabetes or steroid-induced diabetes during the measurement year or the year prior to the measurement year.

Step 3. Determine the numerator: the number of patients who have one or more HbA1c tests and one or more LDL-C tests performed during the measurement year.
Step 4. Calculate the 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.
N/A

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

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.
This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via NCQA's online data submission system.

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)*
Health Plan, Integrated Delivery System, 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
1934-SMD-Testing_Form_v7.1_FINAL.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.
Yes

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.
Yes

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.
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM—Measure Testing

Measure Number (*if previously endorsed*)**:** 1934
Measure Title: Diabetes monitoring for people with diabetes and schizophrenia (SMD)
Date of Submission: 1/5/2018
Type of Measure:

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

Instructions

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

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

2018 Submission

N/A

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

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.20***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice

☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

2018 Submission

Population for measure score reliability testing: The measure score reliability was calculated from HEDIS data that included 151 Medicaid plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Population for Construct Validity Testing: Construct validity was calculated from HEDIS data that included 145 Medicaid health plans. The measured entities included all Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

2012 Submission

Using Medicaid Analytic Extract (MAX) claims data from 2007 we included beneficiaries from 22 states who met the following criteria (1) enrolled in fee-for-service plans* (2) disability as the basis of eligibility; and (3) continuously enrolled in Medicaid for 10 months.

Data from the following states were included in the analytic samples: Alabama, Alaska, California, Connecticut, DC, Georgia, Idaho, Illinois, Indiana, Iowa, Louisiana, Maryland, Missouri, Mississippi, Nevada, New Hampshire, North Carolina, North Dakota, Oklahoma, South Dakota, 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*)

2018 Submission

Patient sample for measure score reliability testing: In 2016, HEDIS measures covered 47 million Medicaid beneficiaries. Data are summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of Plans	Median number of eligible patients per plan
Medicaid	151	159

Beneficiary Sample for Construct Validity Testing: In 2016, HEDIS measures covered 47 million Medicaid beneficiaries. Data is summarized at the health plan level. Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

Product Type	Number of plans	Median number of eligible patients per plan
Medicaid	151	159

2012 Submission

From the beneficiaries, we drew two analytic samples. Beneficiaries who had a primary diagnosis of schizophrenia on either one inpatient or two outpatient claims on different days were included in our schizophrenia sample. Overall, there were 98,412 beneficiaries in the schizophrenia sample.

Beneficiaries ranged in age from 25 – 64 years. Just under half of the schizophrenia population was female (49.2%). About 7% and 34% of the sample was Hispanic and African-American, respectively.

(*Beneficiaries enrolled in managed care plans (e.g. BHO or HMO plans) that provided usable claims records were included. About 1% of the schizophrenia sample was enrolled in a BHO (1.4%) and 11.5% were enrolled in an HMO).

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
2018 Submission
N/A

1.8 What were the social risk factors that were available and analyzed? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.
2018 Submission
We did not analyze performance by social risk factors.

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
2018 Submission
Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the

confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

2012 Submission

The relevant unit of analysis for the proposed measures is aggregated state-level performance. Therefore, we conducted an analysis of test-retest reliability for state results to assess the reliability of state-level performance. To assess stability of state-level performance over time, we computed quartiles of performance based on the state distribution for each measure and assigned each state a score reflecting each state's performance relative to other states in the distribution during the measurement year. For example, a state in the top quartile of all states in 2007 for a given measure would be assigned a performance quartile score of '1' for 2007. This method was replicated for each measure. Next, we repeated this method using 2008 claims data and examined stability of performance quartile between 2007 and 2008.

We also report Pearson correlations measuring the association between 2007 and 2008 measure performance for the 16 states with data.

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

2018 Submission

Beta-Binomial Statistic:

Medicaid
0.855

2012 Submission

In general, the measure showed good test-retest reliability. Overall, 9 of 16 states (44%) had no change in performance quartile between 2007 and 2008. State performance was correlated at r=0.45, indicating that 2007 performance on this measure accounted for 21% of the variance in 2008 scores.

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?*)

Interpretation of measure score reliability testing: The testing suggests the measure has strong reliability.

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

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

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

2018 Submission

We assessed construct and face validity for this measure.

Method of testing construct validity: We tested for construct validity by exploring whether the Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications measure is correlated with the Diabetes Monitoring for People With Diabetes and Schizophrenia measure. We hypothesized that organizations that perform well on Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who are Using Antipsychotic Medications should perform well on Diabetes Monitoring for People With Diabetes and Schizophrenia because the two measures both focus on patients with schizophrenia and whether they received care for diabetes.

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

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

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

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

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. On average, NCQA receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM and NCQA's Board of Directors will be included in the next HEDIS year and reported as first-year measures.

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

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

STEP 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review, and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation, information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.
Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the new year's HEDIS Volume 2.

2012 Submission
Validity was assessed using several complementary methods.

Face validity was assessed through a multistakeholder Technical Advisory Group responsible for overseeing measure development. Additionally, face validity was captured through a public comment period and a series of focus groups involving the Medicaid Medical Directors Learning Network, Managed Behavioral Health Care Organizations, and State Mental Health Commissioners and Medical Directors. The panelists assessed the usability and feasibility of the measures.

Concurrent validity was assessed via Medicaid resource utilization from the Medicaid claims data. We examined rates of schizophrenia-related hospital and emergency room utilization as well as total Medicaid costs comparing beneficiaries in the highest and lowest performance quartiles for each measure.

Convergent and discriminant validity were assessed using the Medicaid Analytic Extract (MAX) from Medicaid claims in using 2007 data. Pearson correlation coefficients were used to assess measure correlations. We hypothesized similar measures (e.g. screening and monitoring) would be correlated and (b) process measures would have negative correlations with measures of adverse events (e.g. mental health emergency room utilization).

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

Statistical results of construct validity testing: The results in Table 1 indicate that there is a strong, positive relationship between the Diabetes Monitoring for People with Diabetes and Schizophrenia measure and Cardiovascular Monitoring for People with Cardiovascular Disease. The relationships are statistically significant ($p<0.05$).

Table 1. Correlations in Medicaid Measures – 2016

	Pearson Correlation Coefficient
	Cardiovascular monitoring for people with cardiovascular disease and schizophrenia
Diabetes monitoring for people with diabetes and schizophrenia	0.66

Note: $p<0.05$

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

2012 Submission:
Face validity:
The measures were deemed important, usable, and feasible to collect by the Technical Advisory Group overseeing the measure development, as well as focus groups with the Medicaid Medical Directors Learning Network, Managed Behavioral Healthcare Organizations, and State Mental Health Commissioners and Medical Directors.

Among 22 states, the measure had a minimum value of 9.1%, mean=57.3%, 25th percentile=55.6%, median=62.1%, 75th percentile=67.7% and a maximum value of 81.6%.

Concurrent validity:
Beneficiaries in the lowest performing states the measure had higher rates of schizophrenia related hospitalization and ED use (23.7% and 26.7%, respectively) than individuals in the highest performing states (14.3% and 24.2%, respectively).

Concurrent and discriminant validity:
Performance on the measure was significantly correlated with the cardiovascular screening and monitoring measures (r=0.908 and r =.888, respectively).

2b1.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
2018 Submission
Interpretation of construct validity testing: The two measures had positive and statistically significant correlation, which indicates the measure has good construct validity.

Interpretation of systematic assessment of face validity: NCQA's expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that *Diabetes monitoring for people with diabetes and schizophrenia (SMD)* is measuring what it intends to measure and that the results of the

measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

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*)
Testing was not performed for exclusions.

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

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

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

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)*

2018 Submission

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

2012 Submission

Pearson correlations, means and percentiles are reported.

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

2018 Submission

HEDIS 2017 Variation in Performance across Health Plans

	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Medicaid	278	69.7	7.9	59.6	64.4	70.1	75.3	78.8	10.9	<0.05

EP: Eligible Population, the average denominator size across plans submitting to HEDIS
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

2012 Submission

Among 22 states, the measure had a minimum value of 9.1%, mean=57.3%, 25th percentile=55.6%, median=62.1%, 75th percentile=67.7% and a maximum value of 81.6%.

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

2018 Submission

The difference between the 25th and 75th percentile is statistically significant for the Medicaid product line. For Medicaid plans, there is a 10.9 percentage point gap between 25th and 75th percentile plans. This gap represents an average 30 more patients with schizophrenia or bipolar disorder having both an LDL-C test and an HbA1c test during the measurement year in high performing Medicaid plans compared to low performing plans (estimated from average health plan eligible population).

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

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

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

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

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

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

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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

2018 Submission

This measure is collected with a complete sample.

2012 Submission

There is no bias on this measure due to missing data.

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)
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.
NCQA recognizes that, despite the clear specifications defined for HEDIS measures, data collection and calculation methods may vary, and other errors may taint the results, diminishing the usefulness of HEDIS data for managed care organization (MCO) comparison. In order for HEDIS to reach its full potential, NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO´s ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable "apples-to-apples" comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:
1) information practices and control procedures
2) sampling methods and procedures

3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system is vital to the regular re-evaluation of NCQA measures.

Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the scientific soundness and feasibility of the measure.

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)*.
Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Specific Plan for Use	Current Use (for current use provide URL)
Not in use	Public Reporting Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Health Plan Ratings https://reportcards.ncqa.org/#/health-plans/list Payment Program Physician Value-Based Payment Modifier (VBM) https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html Quality Improvement (external benchmarking to organizations) Physician Feedback/Quality and Resource Use Reports (QRUR) https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/downloads/QRUR_Presentation.pdf Annual State of Health Care Quality

	http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass

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

PHYSICIAN VALUE-BASED PAYMENT MODIFIER (VBM): This measure is used in the Physician Value-Based Modifier which provides differential payment to a physician or group of physicians under the Medicare Physician Fee Schedule (PFSS). VBM is based on the quality of care provided in comparison to the cost of care within a performance period. The Value Modifier is an adjustment made to Medicare payments for items and services under the Medicare PFS.

NCQA STATE OF HEALTH CARE QUALITY REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. In 2017, the report included results from calendar year 2016 for health plans covering over 171 million people.

NCQA HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2016, a total of 472 Medicare Advantage health plans, 413 commercial health plans and 270 Medicaid health plans across 50 states were included in the rankings.

NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.

PHYSICIAN FEEDBACK/QUALITY AND RESOURCE USE REPORTS (QRUR): This measure is used in the Physician Feedback Program and Quality and Resource Use Reports which provide comparative performance information to
Medicare Fee-For-Service physicians. The Quality and Resource Use Reports show physicians the portion of their Medicare fee-for-service (FFS) patients who have received indicated clinical services, how patients utilized services, and how Medicare spending for their patients compares to average Medicare spending.

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.
Health plans that report HEDIS calculate their rates and know their performance when submitting to NCQA. NCQA publicly reports rates across all plans and also creates benchmarks in order to help plans understand how they perform relative to other plans. Public reporting and benchmarking are effective quality improvement methods.

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.**
NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System.

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.
NCQA measures are evaluated regularly. During this "reevaluation" process, we seek broad input on the measure, including input on performance and implementation experience. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

4a2.2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have not reported significant barriers to implementing this measure, as it uses the administrative data collection method. Questions have generally centered around minor clarification of the specifications, such as confirmation that patients are correctly excluded from the measure according to the measure specification. NCQA responded to all questions to ensure consistent implementation of the specifications.

4a2.2.3. **Summarize the feedback obtained from other users**
This measure has been deemed a priority measure by NCQA and other entities.

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.**
Feedback has not required modification to this measure.

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.
From 2015 to 2017, performance rates for this measure have been generally stable or shown slight improvement. In 2017, Medicaid plans had an average performance rate of 70 percent. There continues to be significant variation between the 10th and 90th percentiles, suggesting room for improvement. In 2017, Medicaid plans in the 10th percentile had a rate of 60 percent, compared to 79 percent among plans in the 90th percentile.

This measure was first introduced in HEDIS 2013. Rates for Medicaid were 67.8 percent. In the last 6 years, we have seen improvement of two percent.

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 were no identified unintended consequences for this measure during testing or since implementation.

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

There were no identified unintended consequences for this measure during testing or since implementation.

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)**
1932 : Diabetes Screening for People With Schizophrenia or Bipolar Disorder Who Are Using Antipsychotic Medications (SSD)
1933 : Cardiovascular Monitoring for People With Cardiovascular Disease and Schizophrenia (SMC)

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

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.**
N/A

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.)
N/A

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:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** National Committee for Quality Assurance

Co.2 **Point of Contact:** Bob, Rehm, nqf@ncqa.org, 202-955-1728-
Co.3 **Measure Developer if different from Measure Steward:** National Committee for Quality Assurance
Co.4 **Point of Contact:** Kristen, Swift, Swift@ncqa.org, 202-955-5174-

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 Technical Advisory Group advised Mathematica Policy Research, Inc. and the National Committee for Quality Assurance during measure development. The TAG was responsible for providing feedback on measure concepts, specifications, results from field and data testing. The TAG consisted of a multistakeholder group of experts with knowledge in behavioral health and quality measurement.

Technical Advisory Group Roster:
Alisa Busch, MD, MS
Enola Proctor, PhD, MSW
David Shern, PhD
Wilma Townsend, MSW
Dan Ford, MD, MPH
Lorrie Rickman-Jones, PhD
Eric Hamilton
Alexander Young, MD, MHS
Peter Delany, PhD
Ben Druss, MD, MPH
Maureen Corcoran, MSN, MBA
Mike Fitzpatrick, MSW
Anita Yuskauskas
Bob Heinssen, PhD

Consultants:
Lisa Dixon, MD, MPH
Julie Kreyenbul, PharmD, PhD

COMMITTEE ON PERFORMANCE MEASUREMENT:
Bruce Bagley, MD, FAAFP, Independent Consultant
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Siemens Healthineers
Helen Darling, MA, Strategic Advisor on Health Benefits & Health Care
Andrea Gelzer, MD, MS, FACP, AmeriHealth Caritas
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Washington Permanente Medical Group
Christine Hunter, MD, (Co-Chair) US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services
Nancy Lane, PhD, Independent Consultant
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MBA, Montefiore Health System
Wayne Rawlins, MD, MBA, ConnectiCare
Rodolfo Saenz, MD, MMM, FACOG, Riverside Medical Clinic
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Adaptive Health
JoAnn Volk, MA, Reforms
Lina Walker, PhD, AARP

Behavioral Health Measurement Advisory Panel:
Katharine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute
Christopher Dennis, MD, MBA, FAPA, Landmark Health, LLC

Ben Druss, MD, MPH, Emory University
Frank Ghinassi, PhD, ABPP, Rutgers University Behavioral Health Care
Connie Horgan, ScD, Brandeis University
Laura Jacobus-Kantor, PhD, SAMHSA
Jeffrey Meyerhoff, MD, Optum
Harold Pincus, MD, College of Physicians and Surgeons, Columbia University, New York Presbyterian Hospital, RAND
Michael Schoenbaum, PhD, National Institute of Mental Health
John Straus, MD, Massachusetts Behavioral Health Partnership-A Beacon Health Options Company

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2012

Ad.3 **Month and Year of most recent revision:** 04, 2018

Ad.4 **What is your frequency for review/update of this measure?** Evrey 3-5 years.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2019

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 "AS IS" WITHOUT WARRANTY OF ANY KIND.

Ad.8 **Additional Information/Comments:** NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties, or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures, and shall not disassemble, recompile, or reverse engineer the source code or object code relating to the measures. Anyone desiring to use or reproduce the measures without modification for a noncommercial purpose may do so without obtaining approval from NCQA. All commercial uses must be approved by NCQA and are subject to a license at the discretion of NCQA

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 3389
Measure Title: Concurrent Use of Opioids and Benzodiazepines (COB)
Measure Steward: PQA, Inc.
Brief Description of Measure: The percentage of individuals 18 years and older with concurrent use of prescription opioids and benzodiazepines during the measurement year.

A lower rate indicates better performance.
Developer Rationale: Overdose deaths involving prescription opioids were five times higher in 2106 than in 1999, and more than 200,000 people have died in the U.S. from overdoses related to prescription opioids.(1,2) Scientific research has identified high-risk prescribing practices that have contributed to the opioid overdose epidemic, including overlapping opioid and benzodiazepine prescriptions.(3) Concurrent use of opioids and benzodiazepines, both central nervous system (CNS) depressants, increases the risk for severe respiratory depression, which can be fatal.(3,4)

According to the Centers for Disease Control and Prevention (CDC) Guideline for Prescribing Opioids for Chronic Pain – United States, 2016, clinicians should avoid concurrent prescribing of opioids and benzodiazepines whenever possible.(3) This is a Category A recommendation (applies to all persons; most patients should receive the recommended course of action) and is based on Type 3 evidence (observational studies or randomized clinical trials with notable limitations). In August 2016, the US Food and Drug Administration added concurrent use of opioids and benzodiazepines as a black box warning to prescription opioids (analgesic and cough medicine) and benzodiazepines.(4)

Several studies indicate that concurrent use of opioids and benzodiazepines puts patients at greater risk for a fatal overdose. Three studies of opioid overdose deaths found evidence of concurrent benzodiazepine use in 31%–61% of cases.(5-7) In the United States, the number of opioid overdose deaths involving benzodiazepines increased 14% on average for each year from 2006 through 2011. However, the number of opioid overdose deaths not involving benzodiazepines did not change significantly.(8) A case-cohort study found that concurrent use of benzodiazepines among US veterans raised the risk of drug overdose deaths four-fold (hazard ratio, 3.86, 95% confidence interval [CI] 3.49-4.26) compared with patients not using benzodiazepines.(9) In a large sample of privately insured patients from 2001-2013, opioid users who also used benzodiazepines were at substantially higher risk of an emergency department (ED) visit or hospital admission for opioid overdose (adjusted odds ratio 2.14; 95% CI, 2.05-2.24). If this association is causal, elimination of the concurrent use could reduce the population risk of an ED visit or hospitalization for opioid overdose by 15%.(10)

Despite the risks, concurrent prescriptions for opioids and benzodiazepines are common and increasing. From 2001-2013, concurrent prescribing (overlap of at least one day) increased by nearly 80% (from 9% to 17%) among privately insured patients.(10) In one study, approximately half of the patients received both opioid and benzodiazepine prescriptions from the same prescriber on the same day.(11) In a 2015 analysis of Medicare Part D non-cancer and/or non-hospice patients on opioid therapy, the prevalence of benzodiazepine concurrent use was 24%.(12)

The PQA Concurrent Use of Opioids and Benzodiazepines measure evaluates a process that correlates with increased risk of opioid overdose. Efforts to prevent opioid overdose deaths should include a multi-faceted approach, including strategies that focus on monitoring and reducing opioid prescribing that has an unfavorable balance of benefit and harm for most patient populations. The measure excludes patients with cancer and those in hospice due to the unique therapeutic goals, ethical considerations, increased opportunities for medical supervision, and balance of risks and benefits with opioid therapy.(3)

1. Hedegaard H, Warner M, Miniño AM. Drug overdose deaths in the United States, 1999–2016. NCHS Data Brief, no 294. Hyattsville, MD: National Center for Health Statistics. 2017/ CDC. Wide-ranging online data for epidemiologic research (WONDER). Atlanta, GA: CDC, National Center for Health Statistics; 2016. Available at http://wonder.cdc.gov
2. Frenk SM, Porter KS, Paulozzi LJ. Prescription opioid analgesic use among adults: United States, 1999–2012. NCHS data brief, no 189. Hyattsville, MD: National Center for Health Statistics. 2015.
3. Dowell D, Haegerich TM, Chou R. CDC Guideline for Prescribing Opioids for Chronic Pain - United States, 2016. MMWR Recomm Rep. 2016;65(1):1-49. doi:10.15585/mmwr.rr6501e1.
4. US Food and Drug Administration. FDA Drug Safety Communication: FDA warns about serious risks and death when combining opioid pain or cough medicines with benzodiazepines; requires its strongest warning. August 31, 2016. Available at: http://www.fda.gov/Drugs/DrugSafety/ucm518473.htm. Accessed: November 9, 2016.
5. Gomes T, Mamdani MM, Dhalla I a, Paterson JM, Juurlink DN. Opioid dose and drug-related mortality in patients with nonmalignant pain. Arch Intern Med. 2011;171(7):686-691. doi:10.1001/archinternmed.2011.117.
6. Dasgupta N, Funk MJ, Proescholdbell S, Hirsch A, Ribisl KM, Marshall S. Cohort Study of the Impact of High-dose Opioid Analgesics on Overdose Mortality. Pain Med. September 2015. doi:10.1111/pme.12907.
7. Jones CM, McAninch JK. Emergency Department Visits and Overdose Deaths From Combined Use of Opioids and Benzodiazepines. Am J Prev Med. 2015;49(4):493-501. doi:10.1016/j.amepre.2015.03.040.
8. Chen LH, Hedegaard H, Warner M. Drug-poisoning Deaths Involving Opioid Analgesics: United States, 1999-2011. NCHS Data Brief. 2014;(166):1-8.
9. Park TW, Saitz R, Ganoczy D, Ilgen MA, Bohnert ASB. Benzodiazepine prescribing patterns and deaths from drug overdose among US veterans receiving opioid analgesics?: case-cohort study. :1-8. doi:10.1136/bmj.h2698.
10. Sun EC, Dixit A, Humphreys K, et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. BMJ. 2017;356:j760. doi: 10.1136/bmj.j760. PMID: 28292769
11. Hwang CS, Kang EM, Kornegay CJ, Staffa JA, Jones CM, McAninch JK. Trends in the Concomitant Prescribing of Opioids and Benzodiazepines, 2002-2014. Am J Prev Med. 2016:1-10. doi:10.1016/j.amepre.2016.02.014.
12. CMS. Concurrent Use of Opioids and Benzodiazepines in a Medicare Part D Population. May 12, 2016. 2016. https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Concurrent-Use-of-Opioids-and-Benzodiazepines-in-a-Medicare-Part-D-Population-CY-2015.pdf. Accessed December 6, 2016.

Numerator Statement: The number of individuals from the denominator with concurrent use of opioids and benzodiazepines for 30 or more cumulative days during the measurement year.

Denominator Statement: The denominator includes individuals 18 years and older with 2 or more prescription claims for opioids with unique dates of service, for which the sum of the days' supply is 15 or more days. Individuals with cancer or in hospice are excluded.

Denominator Exclusions: Individuals with cancer or in hospice at any point during the measurement year are excluded from the denominator.

Measure Type: Process
Data Source: Claims
Level of Analysis: Health Plan

New Measure - Preliminary Analysis

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 describing the lack of therapeutic benefit and increased risk for overdoes as the rationale to support the measure of concurrent use of opioids and benzodiazepines.
- Evidence includes Guidelines and several large scale observational studies:
    - Centers for Disease Control and Prevention (CDC) (2016). Guideline for Prescribing Opioids for Chronic Pain. Recommendation 11 "avoidance of prescribing opioid pain medication and benzodiazepines concurrently". **Grade: recommendation category A (applies to all persons) and evidence type 3 (observational studies or RTC with notable limitations).**
    - Sun et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis (2017). (N=315,426)
    - Gaither et al. The Association Between Receipt of Guideline-Concordant Long-Term Opioid Therapy and All-Cause Mortality (2016) (N=17,044)
    - Dasgupta et al. Cohort Study of the Impact of High-Dose Opioid Analgesics on Overdose Mortality (2016) (N=2,182,374)
- In addition, the developer cites the US Food and Drug Administration (FDA) Boxed Warnings for prescription opioid pain and benzodiazepines.

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 guideline and empirical evidence (Box 3) -> QQC presented (Box 4) -> Moderate (Box 5b) -> 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.

- In a 2015 analysis of Medicare Part D non-cancer and/or non-hospice patients on opioid therapy, the prevalence of benzodiazepine concurrent use was 24%.
- The measure was tested on Medicare (5% national sample from 2015) and Medicaid (Medicaid Analytic eXtract data) health plan data sources. Measure rates for both populations include significant variation.

	Min	Max	Mean	Standard Deviation
Medicare	2.1%	44.7%	22.2%	7.3%
Medicaid	0.0%	17.3%	5.0%	3.5%

Disparities

- The beneficiary level Low-Income Subsidy (LIS) variable was used to determine disparities in rates for populations with different sociodemographic status. The LIS is a subsidy paid by the Federal government to the drug plan for Medicare beneficiaries who need extra help with their prescription drug costs due to limited income and resources. The measure rate for the LIS group was 29.9% while the rate for the non-LIS population was significantly lower, at 19.9%.

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:

**Yes there is evidence

1b. Performance Gap

Comments:

**Yes there is a gap.

**Yes fairly large gap.

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)?*
- *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:

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

Committee pre-evaluation comments

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

2a1. Reliability – Specifications
Comments:
**Would be reliable via medication claims.

2a2. Reliability – Testing
Comments:
**No comments received.

2b1. Validity –Testing
2b4-7. Threats to Validity
2b4. Meaningful Differences
Comments:
**No comments received.

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:
**No issues re: validity.
**I suspect the major threat is missing data. As more and more insurance companies put edits in place for opioids and benzos you will find more members paying cash. This does not take away the risk, but does create missing data.

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.

- Pilot sites testing measure indicated that measure was feasible and results were reported efficiently, accurately, and without difficulty.
- The required data (prescription and medical claims) are readily available in electronic format.
- Measure developer (PQA) retains the rights to measure and can rescind or alter the measure at any time.

Questions for the Committee:

- o *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:
**Using medication data this is very feasible I think.

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**

Accountability program details

- Measure was added in 2018 to Centers for Medicare & Medicaid Services (CMS) Medicaid Adult Core Measure Set. CMS annually releases information on state progress in reporting the Adult Core Set measures and assesses state-specific performance for measures that are reported by at least 25 states and which met internal standards of data quality.

- This is the first year the measure has been included in core set, and was not included in public reporting. Developer anticipates adoption of the measure over time to 25 state threshold for public reporting.

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

Additional Feedback:

- N/A

***Questions for the Committee**:*

- *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.

- The performance results can be used to establish benchmarks and identify opportunities to decrease co-prescribing of opioid and benzodiazepines.

Improvement results

- N/A

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 identified.

Potential harms

- None identified.

Additional Feedback:

- N/A

***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**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

**No comments received.

4b1. Usability – Improvement

Comments:

** Potentially incompletely treated pain or anxiety. Especially on a temporary basis. Could consider adding a dose for the measure. I would not lower the measure to decrease days below the 15. Benzo and opioids are commonly used together in outpatient surgery/procedure setting where the member get the medication and takes it at the center. This very common practice for cosmetic centers

Criterion 5: Related and Competing Measures

Related or competing measures

Related measures include:

- NQF #2940 : Use of Opioids at High Dosage in Persons Without Cancer
- NQF #2950 : Use of Opioids from Multiple Providers in Persons Without Cancer
- NQF #2951 : Use of Opioids from Multiple Providers and at High Dosage in Persons Without Cancer
- Use of Opioids at High Dosage (NCQA)
- Use of Opioids from Multiple Providers (NCQA)

Harmonization

- The PQA opioid measures (NQF # 2940, 2950, and 2951) use the same target population (denominator), and each have different areas of focus (numerator) related to opioid prescribing. The NCQA opioid measures were developed as an adaptation to existing PQA measures; the NCQA opioid measure denominators are similar to the PQA opioid measures, but have a different area of focus than the concurrent use of opioids and benzodiazepines measure.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 3389
Measure Title: Concurrent Use of Opioids and Benzodiazepines (COB)

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document

   ***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?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐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, skip Questions #3-8, then go to Question #9)

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

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)

   Beta-binomial model was used to calculate signal-to-noise ratio of computed measure scores for individual plans reliability scores.

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #6)

   ☒Moderate (go to Question #6)

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

   ☐Insufficient (go to Question #6)

   Reliability results on Medicare (pt. D) and Medicaid MAX data .773 and .937 for MAX Data (Reliability is stronger in Medicaid MAX vs. Medicare Part D)

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

   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*

   ☐Yes (go to Question #7)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   REFERENCE: Testing attachment, section 2a2.2
   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 #8)
   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?
   REFERENCE: Testing attachment, section 2a2
   TIPS*: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)
   TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but* ***check with NQF staff before proceeding, to verify.***
   ☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)
   ☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?
    REFERENCE: Testing attachment, section 2b2-2b6
    TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*
    ☒Yes (go to Question #12)
    ☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

    No concerns were identified, however, the Medicare population has a higher hospice exclusion rate.
12. Analysis of potential threats to validity: Any concerns with measure exclusions?
    REFERENCE: Testing attachment, section 2b2.
    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 #13)
    ☒No (go to Question #13)
    ☐Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to <u>all</u> outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)
    REFERENCE: Testing attachment, section 2b3.

    13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

    13b. Are social risk factors included in risk model? ☐Yes ☐No

    13c. Any concerns regarding the risk-adjustment approach?
    TIPS*: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*
    ☐Yes (please explain below then go to Question #14)
    ☐No (go to Question #14)
    ☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

    N/A Process measure

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
    REFERENCE: Testing attachment, section 2b4.
    ☐Yes (please explain below then go to Question #15)
    ☒No (go to Question #15)

For Medicare population the mean rate was 22.2%, with a median rate of 21.4%, with the lowest plan contract rate at 2.1% and the highest plan contract rate of 44.7%.

For Medicaid MAX population the mean rate was 5.0%, with a median rate of 4.5%. The lowest plan contract rate was 0.0% and the highest plan contract rate was 17.3%.

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.
☐Yes (please explain below then go to Question #16)
☒No (go to Question #16)
☐Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.
☐Yes (please explain below then go to Question #17)
☒No (go to Question #17)

Assessment of Measure Testing

17. Was <u>empirical</u> validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: only face validity testing was performed; 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 #18)
☒No (please explain below, then skip Questions #18-23 and go to Question #24)
New measure – Face validity.

18. Was validity testing conducted with <u>computed performance measure scores</u> for each measured entity?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Answer no if: one overall score for all patients in sample used for testing patient-level data.*
☐Yes (go to Question #19)
☐No (please explain below, then skip questions #19-20 and go to Question #21)

19. Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?
REFERENCE: Testing attachment, section 2b1.
***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 #20)
☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

20. **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 #21)
☐Moderate (go to Question #21)
☐Low (please explain below then go to Question #21)
☐Insufficient (go to Question #21)

21. Was validity testing conducted with patient-level data elements?
REFERENCE: Testing attachment, section 2b1.
***TIPS**: Prior validity studies of the same data elements may be submitted*
☐Yes (go to Question #22)
☐No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*
***REFERENCE:** Testing attachment, section 2b1.*
***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 #23)
☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)
☐Low (please explain below, skip Questions #24-25 and go to Question #26)
☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?
NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]
REFERENCE: Testing attachment, section 2b1.
***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 #25)
☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?
REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☒Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **OVERALL RATING OF VALIDITY** taking into account the results and scope of <u>all</u> 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 <u>are</u> threats to validity and/or threats to validity were <u>not assessed</u>]

☐Insufficient (if insufficient, please explain below) [NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level <u>is required</u>; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

Out of the 16 members of the TEP who voted on the measure, 93.8% recommended that the draft measure be considered for endorsement by the PQA membership, considering the criteria of importance, scientific acceptability, feasibility, and usefulness.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Concurrent Use of Opioids and Benzodiazepines
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: 3/30/2018

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: Concurrent use of opioid medications and benzodiazepine medications

☐ 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.

The measured process, concurrent use of opioids and benzodiazepines, correlates with negative health outcomes. Scientific research has identified high-risk prescribing practices that have contributed to the opioid overdose epidemic, including overlapping opioid and benzodiazepine prescriptions. The *Centers for Disease Control and Prevention (CDC) Guideline for Prescribing Opioids for Chronic Pain – United States, 2016*, provides a category A recommendation (applies to all persons; most patients should receive the recommended course of action) that prescribers should avoid concurrent prescriptions of opioids and benzodiazepines. Opioids and benzodiazepines are both central nervous system (CNS) depressants and can increase the risk for severe respiratory depression and fatal overdose. Few medication situations warrant concurrent use of opioids and benzodiazepines, specifically oncology and hospice, which are excluded from the measure. The lack of a therapeutic benefit combined with increased risk for overdose is the rationale to support this process measure.

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

****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

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**	• CDC Guideline for Prescribing Opioids for Chronic Pain - United States, 2016. • Dowell D, Haegerich TM, Chou R. • March 18, 2016 • MMWR Recomm Rep. 2016 Mar 18;65(1):1-49. doi: 10.15585/mmwr.rr6501e1. • Available at: http://www.cdc.gov/drugoverdose/prescribing/guideline.html • Also, the associated Clinical Evidence Review (http://stacks.cdc.gov/view/cdc/38026), and Contextual Evidence Review (http://stacks.cdc.gov/view/cdc/38027).
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.	CDC Guideline: Recommendation 11, pages 31-32, "Clinicians should avoid prescribing opioid pain medication and benzodiazepines concurrently whenever possible (recommendation category: A, evidence type: 3)."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	CDC Guideline: Type 3 evidence: Observational studies or randomized clinical trials with notable limitations.
Provide all other grades and definitions from the evidence grading system	CDC Guideline: Evidence Type: Based on study design as well as a function of limitations in study design or implementation, imprecision of estimates, variability in findings, indirectness of evidence, publication bias, magnitude of treatment effects, dose-response gradient, and constellation of plausible biases that could change effects. Type 1 evidence: Randomized clinical trials or overwhelming evidence from observational studies. Type 2 evidence: Randomized clinical trials with important limitations, or exceptionally strong evidence from observational studies. Type 3 evidence: Observational studies or randomized clinical trials with notable limitations. Type 4 evidence: Clinical experience and observations, observational studies with important limitations, or randomized clinical trials with several major limitations.
Grade assigned to the **recommendation** with definition of the grade	CDC Guideline: Category A recommendation: Applies to all persons; most patients should receive the recommended course of action.
Provide all other grades and definitions from the recommendation grading system	CDC Guideline: Recommendation Categories Based on evidence type, balance between desirable and undesirable effects, values and preferences, and resource allocation (cost). Category A recommendation: Applies to all persons; most patients should receive the recommended course of action. Category B recommendation: Individual decision making needed;

	different choices will be appropriate for different patients. Clinicians help patients arrive at a decision consistent with patient values and preferences and specific clinical situations.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	• Quantity: four studies • Quality: Observational studies; a) three epidemiologic series of concurrent benzodiazepine use in large proportions of opioid-related overdose deaths, and b) one case-cohort study.
Estimates of benefit and consistency across studies	Not provided.
What harms were identified?	The Clinical Evidence Review did not address risks of benzodiazepine co-prescription among patients prescribed opioids. However, the Contextual Evidence Review found evidence in epidemiologic series of concurrent benzodiazepine use in large proportions of opioid-related overdose deaths, and a case-cohort study found concurrent benzodiazepine prescription with opioid prescription to be associated with a near quadrupling of risk for overdose death compared with opioid prescription alone.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	1. Sun EC, Dixit A, Humphreys K, et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. BMJ. 2017;356:j760. doi: 10.1136/bmj.j760. PMID: 28292769 2. Gaither JR, Goulet JL, Becker WC, et al. The Association Between Receipt of Guideline-Concordant Long-Term Opioid Therapy and All-Cause Mortality. J Gen Intern Med 2016; 31:492 3. Dasgupta N, Funk MJ, Proescholdbell S, et al. Cohort Study of the Impact of High-Dose Opioid Analgesics on Overdose Mortality. Pain Med 2016; 17:85. The studies listed above do not change the conclusion from the SR. All support that the measured process correlates with negative health outcomes.

_____

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.

N/A

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

In a retrospective observational study (N=315,428), Sun et al. reported that opioid users who also used benzodiazepines were at higher risk of an emergency department visit or hospital admission for opioid overdose (adjusted odds ratio [aOR] 2.14; 95% Confidence Interval [CI], 2.05-2.24). The authors estimated that the elimination of the concurrent use of opioids and benzodiazepines could reduce the population risk of an emergency department visit or hospital admission for opioid overdose by 15%.

In a retrospective observational study (N=17,044), Gaither et al. evaluated the association between receipt of guideline-concordant long-term opioid therapy (>90 days) among HIV-infected patients with 1-year all-cause mortality. Patients prescribed benzodiazepines concurrent with opioids, defined as pharmacy documentation that the patient was

prescribed a benzodiazepine greater than 7 days between start date and end of 180 days of long-term opioid therapy, had a higher risk of mortality (matched hazard ratio [HR] 1.39; 95% CI, 1.12-1.66).

In a prospective observational cohort study with one year of follow-up (N=2,182,374 with opioid prescriptions), Dasgupta et al. observed that rates of overdose death among patients on concurrent opioids and benzodiazepines in North Carolina were ten times higher (7 per 10,000 person-years; 95% CI 6.3-7.8) than opioid monotherapy (0.7 per 10,000 person-years; 95% CI 0.6-0.9).

In August 2016, the US Food and Drug Administration (FDA) added Boxed Warnings to prescription drug labeling for prescription opioid pain and prescription opioid cough medications, and benzodiazepines, based on a review of the literature that found that combined use of opioids with benzodiazepines or other drugs that depress the central nervous system (CNS) has resulted in serious side effects, including slowed or difficult breathing and deaths.

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

A primary search of the literature was conducted via PubMed for clinical trials and observational studies (April 2015 through February 2018), and a search of the FDA website was conducted.

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

1. Sun EC, Dixit A, Humphreys K, et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. BMJ. 2017;356:j760. doi: 10.1136/bmj.j760. PMID: 28292769
2. Gaither JR, Goulet JL, Becker WC, et al. The Association Between Receipt of Guideline-Concordant Long-Term Opioid Therapy and All-Cause Mortality. J Gen Intern Med 2016; 31:492
3. Dasgupta N, Funk MJ, Proescholdbell S, et al. Cohort Study of the Impact of High-Dose Opioid Analgesics on Overdose Mortality. Pain Med 2016; 17:85.
4. US Food and Drug Administration. FDA Drug Safety Communication: FDA warns about serious risks and death when combining opioid pain or cough medicines with benzodiazepines; requires its strongest warning. August 31, 2016. Available at: http://www.fda.gov/Drugs/DrugSafety/ucm518473.htm. Accessed: November 9, 2016.

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 3389
Corresponding Measures:
De.2. **Measure Title:** Concurrent Use of Opioids and Benzodiazepines (COB)
Co.1.1. **Measure Steward:** PQA, Inc.
De.3. **Brief Description of Measure:** The percentage of individuals 18 years and older with concurrent use of prescription opioids and benzodiazepines during the measurement year.

A lower rate indicates better performance.
1b.1. **Developer Rationale:** Overdose deaths involving prescription opioids were five times higher in 2106 than in 1999, and more than 200,000 people have died in the U.S. from overdoses related to prescription opioids.(1,2) Scientific research has identified high-risk prescribing practices that have contributed to the opioid overdose epidemic, including overlapping opioid and benzodiazepine prescriptions.(3) Concurrent use of opioids and benzodiazepines, both central nervous system (CNS) depressants, increases the risk for severe respiratory depression, which can be fatal.(3,4)

According to the Centers for Disease Control and Prevention (CDC) Guideline for Prescribing Opioids for Chronic Pain – United States, 2016, clinicians should avoid concurrent prescribing of opioids and benzodiazepines whenever possible.(3) This is a Category A recommendation (applies to all persons; most patients should receive the recommended course of action) and is based on Type 3 evidence (observational studies or randomized clinical trials with notable limitations). In August 2016, the US Food and Drug Administration added concurrent use of opioids and benzodiazepines as a black box warning to prescription opioids (analgesic and cough medicine) and benzodiazepines.(4)

Several studies indicate that concurrent use of opioids and benzodiazepines puts patients at greater risk for a fatal overdose. Three studies of opioid overdose deaths found evidence of concurrent benzodiazepine use in 31%–61% of cases.(5-7) In the United States, the number of opioid overdose deaths involving benzodiazepines increased 14% on average for each year from 2006 through 2011. However, the number of opioid overdose deaths not involving benzodiazepines did not change significantly.(8) A case-cohort study found that concurrent use of benzodiazepines among US veterans raised the risk of drug overdose deaths four-fold (hazard ratio, 3.86, 95% confidence interval [CI] 3.49-4.26) compared with patients not using benzodiazepines.(9) In a large sample of privately insured patients from 2001-2013, opioid users who also used benzodiazepines were at substantially higher risk of an emergency department (ED) visit or hospital admission for opioid overdose (adjusted odds ratio 2.14; 95% CI, 2.05-2.24). If this association is causal, elimination of the concurrent use could reduce the population risk of an ED visit or hospitalization for opioid overdose by 15%.(10)

Despite the risks, concurrent prescriptions for opioids and benzodiazepines are common and increasing. From 2001-2013, concurrent prescribing (overlap of at least one day) increased by nearly 80% (from 9% to 17%) among privately insured patients.(10) In one study, approximately half of the patients received both opioid and benzodiazepine prescriptions from the same prescriber on the same day.(11) In a 2015 analysis of Medicare Part D non-cancer and/or non-hospice patients on opioid therapy, the prevalence of benzodiazepine concurrent use was 24%.(12)

The PQA Concurrent Use of Opioids and Benzodiazepines measure evaluates a process that correlates with increased risk of opioid overdose. Efforts to prevent opioid overdose deaths should include a multi-faceted approach, including strategies that focus on monitoring and reducing opioid prescribing that has an unfavorable balance of benefit and harm for most patient

populations. The measure excludes patients with cancer and those in hospice due to the unique therapeutic goals, ethical considerations, increased opportunities for medical supervision, and balance of risks and benefits with opioid therapy.(3)

1. Hedegaard H, Warner M, Miniño AM. Drug overdose deaths in the United States, 1999–2016. NCHS Data Brief, no 294. Hyattsville, MD: National Center for Health Statistics. 2017/ CDC. Wide-ranging online data for epidemiologic research (WONDER). Atlanta, GA: CDC, National Center for Health Statistics; 2016. Available at http://wonder.cdc.gov
2. Frenk SM, Porter KS, Paulozzi LJ. Prescription opioid analgesic use among adults: United States, 1999–2012. NCHS data brief, no 189. Hyattsville, MD: National Center for Health Statistics. 2015.
3. Dowell D, Haegerich TM, Chou R. CDC Guideline for Prescribing Opioids for Chronic Pain - United States, 2016. MMWR Recomm Rep. 2016;65(1):1-49. doi:10.15585/mmwr.rr6501e1.
4. US Food and Drug Administration. FDA Drug Safety Communication: FDA warns about serious risks and death when combining opioid pain or cough medicines with benzodiazepines; requires its strongest warning. August 31, 2016. Available at: http://www.fda.gov/Drugs/DrugSafety/ucm518473.htm. Accessed: November 9, 2016.
5. Gomes T, Mamdani MM, Dhalla I a, Paterson JM, Juurlink DN. Opioid dose and drug-related mortality in patients with nonmalignant pain. Arch Intern Med. 2011;171(7):686-691. doi:10.1001/archinternmed.2011.117.
6. Dasgupta N, Funk MJ, Proescholdbell S, Hirsch A, Ribisl KM, Marshall S. Cohort Study of the Impact of High-dose Opioid Analgesics on Overdose Mortality. Pain Med. September 2015. doi:10.1111/pme.12907.
7. Jones CM, McAninch JK. Emergency Department Visits and Overdose Deaths From Combined Use of Opioids and Benzodiazepines. Am J Prev Med. 2015;49(4):493-501. doi:10.1016/j.amepre.2015.03.040.
8. Chen LH, Hedegaard H, Warner M. Drug-poisoning Deaths Involving Opioid Analgesics: United States, 1999-2011. NCHS Data Brief. 2014;(166):1-8.
9. Park TW, Saitz R, Ganoczy D, Ilgen MA, Bohnert ASB. Benzodiazepine prescribing patterns and deaths from drug overdose among US veterans receiving opioid analgesics?: case-cohort study. :1-8. doi:10.1136/bmj.h2698.
10. Sun EC, Dixit A, Humphreys K, et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. BMJ. 2017;356:j760. doi: 10.1136/bmj.j760. PMID: 28292769
11. Hwang CS, Kang EM, Kornegay CJ, Staffa JA, Jones CM, McAninch JK. Trends in the Concomitant Prescribing of Opioids and Benzodiazepines, 2002-2014. Am J Prev Med. 2016:1-10. doi:10.1016/j.amepre.2016.02.014.
12. CMS. Concurrent Use of Opioids and Benzodiazepines in a Medicare Part D Population. May 12, 2016. 2016. https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Concurrent-Use-of-Opioids-and-Benzodiazepines-in-a-Medicare-Part-D-Population-CY-2015.pdf. Accessed December 6, 2016.

S.4. **Numerator Statement:** The number of individuals from the denominator with concurrent use of opioids and benzodiazepines for 30 or more cumulative days during the measurement year.
S.6. **Denominator Statement:** The denominator includes individuals 18 years and older with 2 or more prescription claims for opioids with unique dates of service, for which the sum of the days' supply is 15 or more days. Individuals with cancer or in hospice are excluded.
S.8. **Denominator Exclusions:** Individuals with cancer or in hospice at any point during the measurement year are excluded from the denominator.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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_Submission_Form_-_PQA_COB_FV-636579943284279616.docx

1a.1 <u>For Maintenance of Endorsement:</u> 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.

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)*

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

Overdose deaths involving prescription opioids were five times higher in 2106 than in 1999, and more than 200,000 people have died in the U.S. from overdoses related to prescription opioids.(1,2) Scientific research has identified high-risk prescribing practices that have contributed to the opioid overdose epidemic, including overlapping opioid and benzodiazepine prescriptions.(3) Concurrent use of opioids and benzodiazepines, both central nervous system (CNS) depressants, increases the risk for severe respiratory depression, which can be fatal.(3,4)

According to the Centers for Disease Control and Prevention (CDC) Guideline for Prescribing Opioids for Chronic Pain – United States, 2016, clinicians should avoid concurrent prescribing of opioids and benzodiazepines whenever possible.(3) This is a Category A recommendation (applies to all persons; most patients should receive the recommended course of action) and is based on Type 3 evidence (observational studies or randomized clinical trials with notable limitations). In August 2016, the US Food and Drug Administration added concurrent use of opioids and benzodiazepines as a black box warning to prescription opioids (analgesic and cough medicine) and benzodiazepines.(4)

Several studies indicate that concurrent use of opioids and benzodiazepines puts patients at greater risk for a fatal overdose. Three studies of opioid overdose deaths found evidence of concurrent benzodiazepine use in 31%–61% of cases.(5-7) In the United States, the number of opioid overdose deaths involving benzodiazepines increased 14% on average for each year from 2006 through 2011. However, the number of opioid overdose deaths not involving benzodiazepines did not change significantly.(8) A case-cohort study found that concurrent use of benzodiazepines among US veterans raised the risk of drug overdose deaths four-fold (hazard ratio, 3.86, 95% confidence interval [CI] 3.49-4.26) compared with patients not using benzodiazepines.(9) In a large sample of privately insured patients from 2001-2013, opioid users who also used benzodiazepines were at substantially higher risk of an emergency department (ED) visit or hospital admission for opioid overdose (adjusted odds ratio 2.14; 95% CI, 2.05-2.24). If this association is causal, elimination of the concurrent use could reduce the population risk of an ED visit or hospitalization for opioid overdose by 15%.(10)

Despite the risks, concurrent prescriptions for opioids and benzodiazepines are common and increasing. From 2001-2013, concurrent prescribing (overlap of at least one day) increased by nearly 80% (from 9% to 17%) among privately insured patients.(10) In one study, approximately half of the patients received both opioid and benzodiazepine prescriptions from the same prescriber on the same day.(11) In a 2015 analysis of Medicare Part D non-cancer and/or non-hospice patients on opioid therapy, the prevalence of benzodiazepine concurrent use was 24%.(12)

The PQA Concurrent Use of Opioids and Benzodiazepines measure evaluates a process that correlates with increased risk of opioid overdose. Efforts to prevent opioid overdose deaths should include a multi-faceted approach, including strategies that focus on monitoring and reducing opioid prescribing that has an unfavorable balance of benefit and harm for most patient populations. The measure excludes patients with cancer and those in hospice due to the unique therapeutic goals, ethical considerations, increased opportunities for medical supervision, and balance of risks and benefits with opioid therapy.(3)

1. Hedegaard H, Warner M, Miniño AM. Drug overdose deaths in the United States, 1999–2016. NCHS Data Brief, no 294. Hyattsville, MD: National Center for Health Statistics. 2017/ CDC. Wide-ranging online data for epidemiologic research (WONDER). Atlanta, GA: CDC, National Center for Health Statistics; 2016. Available at http://wonder.cdc.gov
2. Frenk SM, Porter KS, Paulozzi LJ. Prescription opioid analgesic use among adults: United States, 1999–2012. NCHS data brief, no 189. Hyattsville, MD: National Center for Health Statistics. 2015.
3. Dowell D, Haegerich TM, Chou R. CDC Guideline for Prescribing Opioids for Chronic Pain - United States, 2016. MMWR Recomm Rep. 2016;65(1):1-49. doi:10.15585/mmwr.rr6501e1.
4. US Food and Drug Administration. FDA Drug Safety Communication: FDA warns about serious risks and death when combining opioid pain or cough medicines with benzodiazepines; requires its strongest warning. August 31, 2016. Available at: http://www.fda.gov/Drugs/DrugSafety/ucm518473.htm. Accessed: November 9, 2016.
5. Gomes T, Mamdani MM, Dhalla I a, Paterson JM, Juurlink DN. Opioid dose and drug-related mortality in patients with nonmalignant pain. Arch Intern Med. 2011;171(7):686-691. doi:10.1001/archinternmed.2011.117.
6. Dasgupta N, Funk MJ, Proescholdbell S, Hirsch A, Ribisl KM, Marshall S. Cohort Study of the Impact of High-dose Opioid Analgesics on Overdose Mortality. Pain Med. September 2015. doi:10.1111/pme.12907.
7. Jones CM, McAninch JK. Emergency Department Visits and Overdose Deaths From Combined Use of Opioids and Benzodiazepines. Am J Prev Med. 2015;49(4):493-501. doi:10.1016/j.amepre.2015.03.040.
8. Chen LH, Hedegaard H, Warner M. Drug-poisoning Deaths Involving Opioid Analgesics: United States, 1999-2011. NCHS Data Brief. 2014;(166):1-8.
9. Park TW, Saitz R, Ganoczy D, Ilgen MA, Bohnert ASB. Benzodiazepine prescribing patterns and deaths from drug overdose among US veterans receiving opioid analgesics?: case-cohort study. :1-8. doi:10.1136/bmj.h2698.
10. Sun EC, Dixit A, Humphreys K, et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. BMJ. 2017;356:j760. doi: 10.1136/bmj.j760. PMID: 28292769
11. Hwang CS, Kang EM, Kornegay CJ, Staffa JA, Jones CM, McAninch JK. Trends in the Concomitant Prescribing of Opioids and Benzodiazepines, 2002-2014. Am J Prev Med. 2016:1-10. doi:10.1016/j.amepre.2016.02.014.
12. CMS. Concurrent Use of Opioids and Benzodiazepines in a Medicare Part D Population. May 12, 2016. 2016. https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Concurrent-Use-of-Opioids-and-Benzodiazepines-in-a-Medicare-Part-D-Population-CY-2015.pdf. Accessed December 6, 2016.

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 different health plan data sources – the Medicare and the Medicaid populations.

For the Medicare population, data used for testing came from the Medicare 5% national sample using data from January 1, 2015 to December 31, 2015. The analysis included 710 Medicare Advantage Prescription Drug plans (MA-PD) and 73 standalone Prescription Drug Plans (PDPs) covering 2,952,360 individuals aged 18 and older.

The Medicare rates ranged from 2.1% (minimum) to 44.7% (maximum). The mean rate was 22.2% with a standard deviation of 7.3%. The 25th percentile was 17.4%, the 50th percentile (median) was 21.4% and the 75th percentile was 27.3%. The interquartile range was 9.9%.

For the Medicaid population, the majority of testing data came from the National Medicaid Analytic eXtract (MAX) data. The data included 322 health plans from 17 states covering 11,745,722 individuals aged 18 and older. In addition, one state Medicaid program with three state-based health plans covering 222,896 individuals 18 years and older was included in the testing using the state's Medicaid administrative claims database.

The Medicaid rates for the national (MAX) data ranged from 0.0% (minimum) to 17.3% (maximum). The mean was 5.0% with a standard deviation of 3.5%. The 25th percentile was 2.4%, the 50th percentile (median) was 4.5% and the 75th percentile was 6.9%. The interquartile range was 4.5%.

For the one state Medicaid program with the three health plans, the rate ranged from 2.8% to 6.3%.

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.*
Disparities data are available for the Medicare population. The testing for the Medicare population came from the Medicare 5% national sample using data from January 1, 2015 to December 31, 2015. The analysis included 710 Medicare Advantage Prescription Drug plans (MA-PD) and 73 standalone Prescription Drug Plans (PDPs) covering 2,952,360 individuals aged 18 and older.

The beneficiary level Low-Income Subsidy (LIS) variable was used to determine disparities in rates for populations with different sociodemographic status. The LIS is a subsidy paid by the Federal government to the drug plan for Medicare beneficiaries who need extra help with their prescription drug costs due to limited income and resources. The measure rate for the LIS group was 29.9% while the rate for the non-LIS population was significantly lower, at 19.9%.

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.)*
https://pqaalliance.org/measures/default.asp

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:** PQA_ICD_Code_Cancer_Value_Set_Feb_2018.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.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) 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).
The number of individuals from the denominator with concurrent use of opioids and benzodiazepines for 30 or more cumulative days during the measurement year.

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 number of individuals from the denominator with:

- 2 or more prescription claims for any benzodiazepine with unique dates of service, AND
- Concurrent use of opioids and benzodiazepines for 30 or more cumulative days.

Complete the steps below to identify individuals with concurrent use of opioids and benzodiazepines:

Step 1: From the denominator population, identify individuals with 2 or more prescriptions claims on unique dates of service for any benzodiazepine (Table COB-B, below) during the measurement year.

Step 2: Of the population identified in Step 1, determine the total days of overlap (concurrent use) between the opioid and benzodiazepine prescriptions during the measurement year.

- Concurrent use is identified using the dates of service and days' supply of an individual's opioid and benzodiazepine prescription drug claims. The days of concurrent use is the sum of the number of days (cumulative) during the measurement year with overlapping days' supply for an opioid and a benzodiazepine. Exclude days of overlap that occur after the end of the measurement year.

Step 3: Count the number of individuals with concurrent use of opioids and benzodiazepines for 30 or more cumulative days. This is the numerator.

Note: When identifying days' supply for opioids (or benzodiazepines):

- Exclude any days' supply that occur after the end of the measurement year.
- Multiple prescription claims with the same date of service: If multiple prescription claims for opioids (or benzodiazepines) are dispensed on the same day, calculate the number of days covered by an opioid using the prescriptions with the longest days' supply.

Table COB-B: Benzodiazepines:

Alprazolam, chlordiazepoxide, clobazam, clonazepam, clorazepate, diazepam, estazolam, flurazepam, lorazepam, midazolam, oxazepam, quazepam, temazepam, triazolam
(note: excludes injectable formulations)

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator includes individuals 18 years and older with 2 or more prescription claims for opioids with unique dates of service, for which the sum of the days' supply is 15 or more days. Individuals with cancer or in hospice are excluded.

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 denominator includes individuals 18 years and older by the first day of the measurement year with 2 or more prescription claims for opioids with unique dates of service, for which the sum of the days' supply is 15 or more days. Use Table COB-A: Opioids, below, to identify the opioid medications for the measure.

Complete the steps below to determine the denominator:

Step 1: Identify individuals aged 18 years and older as of the first day of the measurement year

Step 2: Of those identified in step 1, identify individuals meeting the continuous enrollment criteria.
- To be continuously enrolled, an individual may have no more than one gap in enrollment of up to 31 days during the measurement year. When enrollment is verified monthly, the individual may not have more than a 1-month gap in coverage (i.e., an individual whose coverage lapses for 2 months [60 days] is not considered continuously enrolled).

Step 3: Of those identified in step 2, identify individuals with 2 or more prescription claims for opioids on unique dates of service, for which the sum of the days' supply is 15 or more days' supply during the measurement year.

Step 4: Of those identified in step 3, identify individuals where the earliest prescription for an opioid (i.e. Index Prescription Start Date [IPSD]) is 30 or more days from the last day of the measurement year (January 1 through December 2)

Note: When identifying days' supply for opioids:
- Exclude any days' supply that occur after the end of the measurement year.
- Multiple prescription claims with the same date of service: If multiple prescription claims for opioids are dispensed on the same day, calculate the number of days covered by an opioid using the prescriptions with the longest days' supply.

Table COB-A: Opioids:
buprenorphine, butorphanol, codeine, dihydrocodeine, fentanyl, hydrocodone, hydromorphone, levorphanol, meperidine, methadone, morphine, opium, oxycodone, oxymorphone, pentazocine, tapentadol, tramadol
(note: excludes injectable formulations; includes prescription opioid cough medications; excludes single-agent and combination buprenorphine products used to treat opioid use disorder (i.e., buprenorphine sublingual tablets, Probuphine® Implant kit subcutaneous implant, and all buprenorphine/naloxone combination products).

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Individuals with cancer or in hospice at any point during the measurement year are excluded from the denominator.

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.)*
Hospice exclusion: Exclude any individual in hospice during the measurement year. To identify individuals in hospice:
- Use the hospice indicator from the enrollment database, where available (e.g. Medicare); or
- Use place of service code 34 where a hospice indicator is not available (e.g. Commercial, Medicaid)

Cancer exclusion: Exclude any individuals with cancer during the measurement year. To identify individuals with cancer:

- Using ICD codes, refer to those listed in the file titled, PQA ICD Code Cancer Value Set Feb 2018 and attached in S.2b. The list is based on the American Medical Association-convened Physician Consortium for Performance Improvement Cancer value set (OID: 2.16.840.1.113883.3.526.3.1010). A cancer diagnosis is defined as having at least one claim with any of the listed cancer diagnoses, including primary diagnosis or any other diagnosis fields during the measurement year.
- For Medicare Data, if ICD codes are not available, use Prescription Drug Hierarchical Condition Categories (RxHCCs) 15, 16, 17, 18, 19 for Payment Year 2016 or 2017 to identify cancer exclusions. RxHCCs are available at: https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html

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.)*
The measure is stratified by the following lines of business for the health plan:
- Commercial
- Medicare
- Medicaid

Medicare Plans are further stratified by Low-Income Subsidy (LIS) status.
LIS is a subsidy paid by the Federal government to the drug plan for Medicare beneficiaries who need extra help with their prescription drug costs due to limited income and resources. Medicare beneficiaries apply for the LIS with the Social Security Administration or their State Medicaid agency.

The Medicare Master Beneficiary Summary file contains the Cost Share Group variable used to identify LIS status, which is subsidized Part D coverage. There are 12 monthly variables - where the 01 through 12 at the end of the variable name corresponds with the month (e.g., 01 is January and 12 is December). CMS identifies beneficiaries with fully-subsidized Part D coverage by looking for individuals that have a 01, 02, or 03 for the month. Other beneficiaries who are eligible for the LIS but do not receive a full subsidy have a 04, 05, 06, 07, or 08. The remaining values indicate that the individual is not eligible for subsidized Part D coverage.

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.)*
A. Target population (denominator):

Step 1: Identify individuals aged 18 years and older as of the first day of the measurement year

Step 2: Of those identified in step 1, identify individuals meeting the continuous enrollment criteria.
- To be continuously enrolled, an individual may have no more than one gap in enrollment of up to 31 days during the measurement year. When enrollment is verified monthly, the individual may not have more than a 1-month gap in coverage (i.e., an individual whose coverage lapses for 2 months [60 days] is not considered continuously enrolled).

Step 3: Of those identified in step 2, identify individuals with 2 or more prescription claims for opioids on unique dates of service, for which the sum of the days' supply is 15 or more days' supply during the measurement year.

Step 4: Of those identified in step 3, identify individuals where the earliest prescription for an opioid (i.e. Index Prescription Start Date [IPSD]) is 30 or more days from the last day of the measurement year (January 1 through December 2)

Note: When identifying days' supply for opioids:
- Exclude any days' supply that occur after the end of the measurement year.
- Multiple prescription claims with the same date of service: If multiple prescription claims for opioids are dispensed on the same day, calculate the number of days covered by an opioid using the prescriptions with the longest days' supply.

Step 5: Identify individuals with cancer or in hospice during the measurement year.
To identify individuals in hospice:
- Use the hospice indicator from the enrollment database, where available (e.g. Medicare); or
- Use place of service code 34 where a hospice indicator is not available (e.g. Commercial, Medicaid)

To identify individuals with cancer:
- Using ICD codes, refer to those listed in the file titled, PQA ICD Code Cancer Value Set Feb 2018 and attached in S.2b. The list is based on the American Medical Association-convened Physician Consortium for Performance Improvement Cancer value set (OID: 2.16.840.1.113883.3.526.3.1010). A cancer diagnosis is defined as having at least one claim with any of the listed cancer diagnoses, including primary diagnosis or any other diagnosis fields during the measurement year.
- For Medicare Data, if ICD codes are not available, use Prescription Drug Hierarchical Condition Categories (RxHCCs) 15, 16, 17, 18, 19 for Payment Year 2016 or 2017 to identify cancer exclusions. RxHCCs are available at: https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Risk-Adjustors.html

Step 6: Exclude individuals with cancer or in hospice (Step 5) from those identified in Step 4. This is the denominator.

B. Numerator Population:
Step 7: From the denominator population (from Step 6), identify individuals with 2 or more prescriptions claims on unique dates of service for any benzodiazepine during the measurement year.

Step 8: Of the population identified in Step 7, determine the total days of overlap (concurrent use) between the opioid and benzodiazepine prescriptions during the measurement year.
- Concurrent use is identified using the dates of service and days' supply of an individual's opioid and benzodiazepine prescription drug claims. The days of concurrent use is the sum of the number of days (cumulative) during the measurement year with overlapping days' supply for an opioid and a benzodiazepine. Exclude days of overlap that occur after the end of the measurement year.

Step 9: Count the number of individuals with concurrent use of opioids and benzodiazepines for 30 or more cumulative days. This is the numerator.

Note: When identifying days' supply for opioids (or benzodiazepines):
- Exclude any days' supply that occur after the end of the measurement year.
- Multiple prescription opioid (or benzodiazepine) claims with overlap: For multiple prescription claims for opioids (or benzodiazepines) with overlapping days' supply, count each day in the measurement year only once toward the denominator. There is no adjustment for early fills or overlapping days' supply for opioids (or benzodiazepines).

C. Measure Rate:
Step 10: Divide the number of individuals in the numerator (Step 9) by the denominator (Step 6) and multiply by 100. This is the measure rate reported as a percentage.
- Report the rates separately by line of business (e.g. Medicare, Medicaid, Commercial). For Medicare, report rates for low-income subsidy (LIS) and non-LIS populations separately.

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.
N/A

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

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.
Administrative claims: prescription claims, medical claims, Prescription Drug Hierarchical Condition Categories (RxHCCs)

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)*
Health Plan

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Other
If other: The level of analysis for this measure is the prescription drug health plan, but it contains claims data from multiple care settings, including ambulatory, skilled nursing facility, pharmacy etc.

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_Testing_Form_-_PQA_COB_FV-636579943734123366.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.

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Concurrent Use of Opioids and Benzodiazepines
Date of Submission: 3/30/2018
Type of Measure:

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

Instructions

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

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

The measure was tested in two different health plan data sources – the Medicare and the Medicaid populations.

For the Medicare population, data used for testing came from the Medicare 5% national sample data. The Medicare Part D Prescription Drug Event (PDE) claims were used for the identification of prescription drugs. The 5% medical claims (standard analytic files) were used to identify cancer diagnoses and hospice claims. To identify dates of birth and continuous enrollment, the Medicare Beneficiaries Summary Files (MBSF) were used.

For the Medicaid population, the data used for testing came from Medicaid administrative claims. National Medicaid sample data covering 17 states and 322 health plans were included in the testing using data from the Medicaid Analytic eXtract (MAX) data. In addition, one state Medicaid program with three state-based health plans was included in the testing using the state's Medicaid administrative claims database. Medical claims were used to identify the cancer diagnoses, and the pharmacy claims were used for the identification of prescription drugs.

1.3. What are the dates of the data used in testing? 2014-2016

The testing for the Medicare population used administrative claims data from January 1, 2015 to December 31, 2015. The testing for Medicaid used administrative claims data from January 1, 2014 to December 31, 2014 for the national level MAX dataset, and data from January 1, 2016 to December 31, 2016 for one state-based

Medicaid dataset. The data from these time periods were the most recent, complete, full year data available to testers at the time of testing.

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

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

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

The Medicare testing was conducted using the Medicare 5% sample data – a nationally representative sample, including data from all the states. Of beneficiaries aged 18 years or older by the first day of the measurement year, the data included 710 Medicare Advantage Prescription Drug (MA-PD) plans and 73 standalone Prescription Drug Plans (PDPs). Plans varied in size (see Table 1), with a mean plan size of 2,639 beneficiaries and a median plan size of 353 beneficiaries.

Table 1. Plan Size Distribution for 2015 Medicare Sample

Statistic	Number of Beneficiaries
Mean	2,639
Standard Deviation	14,308
Minimum	1
25th Percentile	44
50th Percentile	353
75th Percentile	1,264
Maximum	228,698
Interquartile Range	1,220

For the Medicaid testing, the national level analysis included 322 health plans covering 17 states with beneficiaries aged 18 years or older. Of the 322 plans, 17 plans were fee-for-service (FFS), and the remaining 305 plans were Medicaid Managed Care plans. There was variation in plan size, with mean plan size of 36,477 beneficiaries, and a median plan size of 2,561 beneficiaries (see Table 2). Fifteen plans were from the South region of the United States (US), 28 plans were from the Northeast region of the US, 173 plans were from states in the Midwest region of the US, and 106 plans were from the West region of the US.

Table 2. Plan Size Distribution for 2014 Medicaid MAX Sample

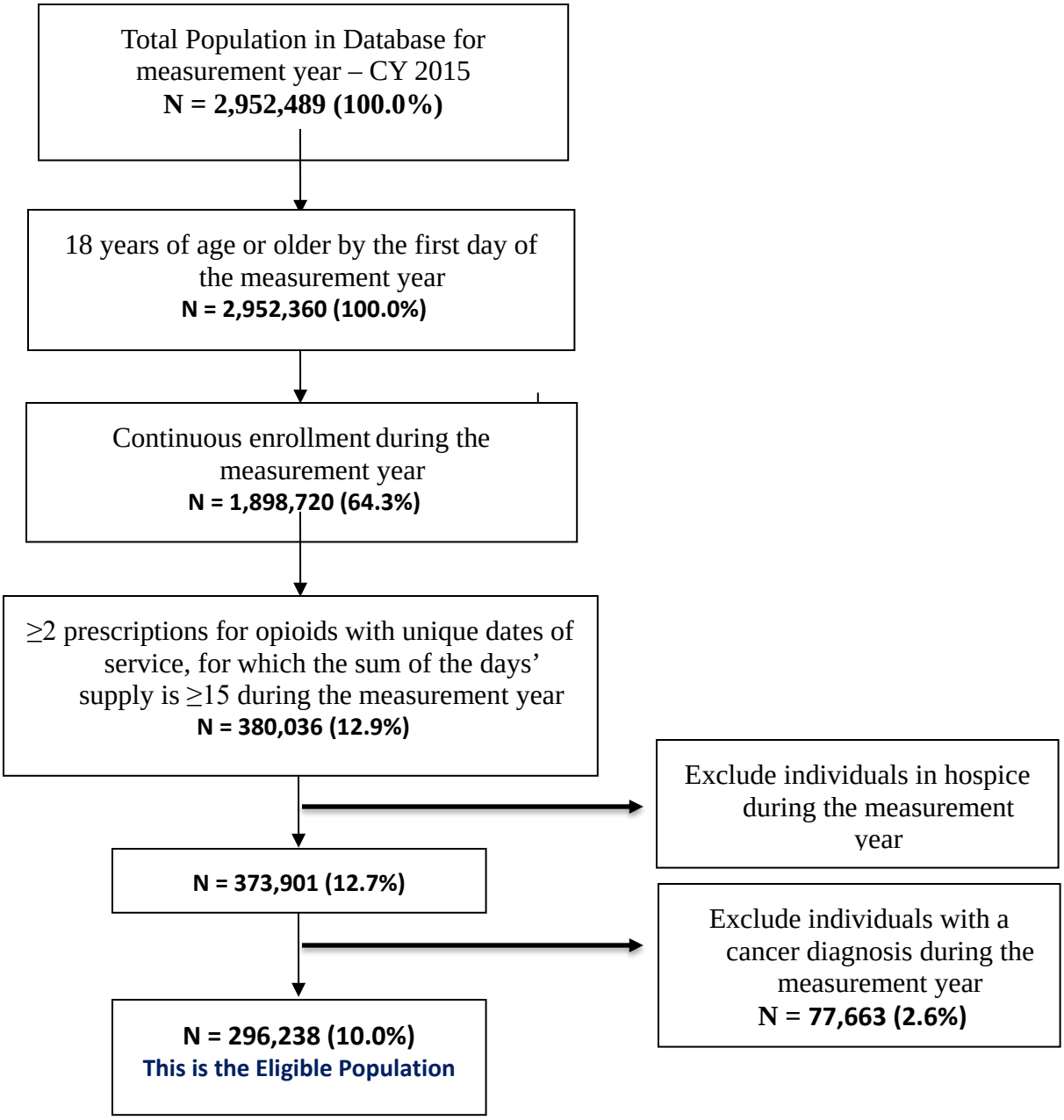
Statistic	Number of Beneficiaries
Mean	36,477
Standard Deviation	185,027
Minimum	2
25th Percentile	101
50th Percentile	2,561
75th Percentile	17,140
Maximum	3,055,163
Interquartile Range	17,039

The one state-based Medicaid program was in the South region and included 3 health plans – 1 FFS and 2 managed care plans. The mean size of the plans was 74,299 beneficiaries.

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*)

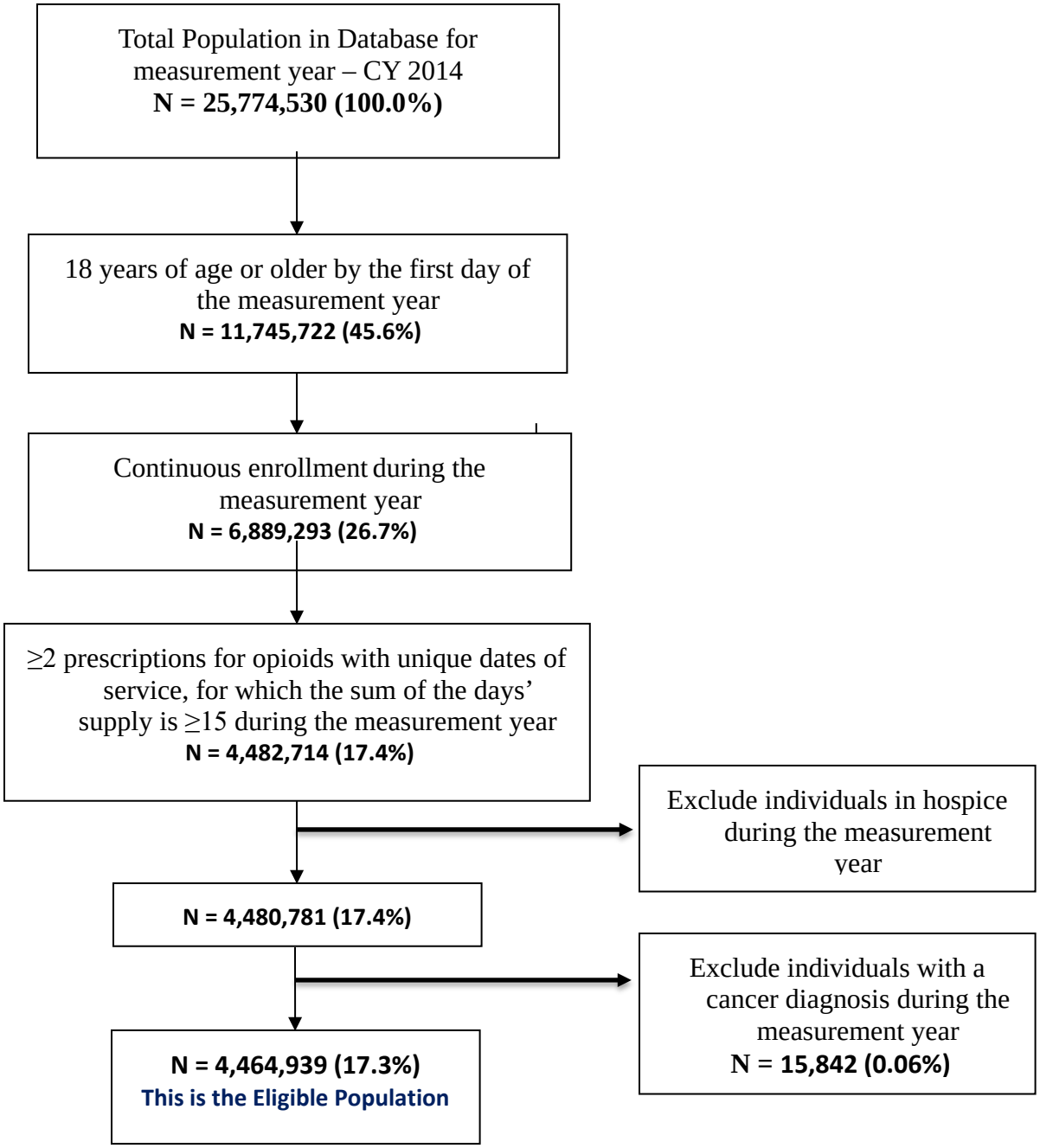
For the Medicare testing, a total of 2,952,360 individuals aged 18 and older were included in the testing and analysis. Of all persons, 1,339,615 (45.4%) were male, and 1,612,745 (54.6%) were female. Individuals by age group included 176,663 (6.0%) age 18 – 50 years, 459,964 (15.6%) age 51 – 64 years, 2,017,849 (68.3%) age 65 – 84 years, and 297,884 (10.1%) age 85 and older. After applying all inclusion and exclusion criteria, the final population for analysis was 296,238 (10.0%) of the initial population. See Figure 1, for the selection criteria for the eligible population for Medicare.

Figure 1. Selection Criteria for Eligible Population for the 2015 Medicare Sample



For the Medicaid MAX population, a total of 11,745,722 beneficiaries age 18 and older were included in the analysis. Of all persons, 4,466,005 (38.0%) were male, and 7,279,717 (62.0%) were female. Individuals by age group included 9,587,267 (81.6%) age 18 – 50 years, 2,016,836 (17.2%) age 51 – 64 years, 129,057 (1.1%) age 65 – 84 years, and 12,562 (0.1%) age 85 and older. After applying all inclusion and exclusion criteria, the final population for analysis was 4,464,939 (17.3%) of the initial population. See Figure 2, for the selection criteria for the eligible population for the Medicaid MAX population.

Figure 2. Selection Criteria for Eligible Population for 2014 Medicaid MAX Sample



Finally, the total population for the 1 state-based Medicaid program was 695,166. Of that initial population, a total of 222,896 beneficiaries age 18 and older were included in the analysis. Of all persons, 53,944 (24.2%) were male, and 168,952 (75.8%) were female. Individuals by age group included 183,647 (82.4%) age 18 – 50 years, 36,535 (16.4%) age 51 – 64 years, 2,614 (1.2%) age 65 – 84 years, and 100 (0.04%) age 85 and older. After applying all inclusion and exclusion criteria, the final population for analysis was 99,390 (14.3%) of the initial population.

As seen in the results above, the measure was tested across a large spectrum of age groups, with the Medicare population being older (primarily 65 years and older), and the Medicaid data included a younger population.

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 testing was conducted for both the Medicare and Medicaid populations. For the Medicare population, reliability testing was conducted at the plan contract level, because the application of this measure in the Medicare program would be assessed at the plan contract level. In accordance with the PQA measure specifications, reliability testing excluded plan contracts with less than 30 individuals in the denominator.

For the Medicaid population, reliability testing was conducted at the plan level using the MAX data, and excluded any plans with less than 30 individuals in the denominator.

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.

For the Medicare population, the beneficiary level Low-Income Subsidy (LIS) variable was used to determine disparities in rates for populations with different sociodemographic status. The LIS is a subsidy paid by the Federal government to the drug plan for Medicare beneficiaries who need extra help with their prescription drug costs due to limited income and resources.

For the Medicaid populations, no patient level indicators of sociodemographic status were available in the data.

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*)

Using the Medicare and Medicaid data described in sections 1.2 to 1.6, the reliability of the computed measure scores was measured as the ratio of signal-to-noise. The signal is the proportion of the variability in measured performance that can be explained by true differences in plan (or contract) performance. Reliability scores range from 0 to 1, with a score of 0 signifying that all variation is due to measurement error. A value of 1 signifies that the variation represents true differences in performance scores between plans. A reliability score of 0.7 is the minimum threshold for reliability.

A beta-binomial model was used to calculate plan-specific reliability scores. This is based on the methods outlined by Adams in the following paper: Adams JL. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND Corporation. 2009. Retrieved from http://www.rand.org/pubs/technical_reports/TR653.

The reliability score is defined as the ratio of the plan-to-plan variance to the sum of the plan-to-plan variance and the plan-specific error. The plan-to-plan variance is an estimate of the variance of the true rates. The plan-specific error variance is the sampling or measurement error.

$$reliability = \frac{\sigma^2_{plan-to-plan}}{\sigma^2_{plan-to-plan} + \sigma^2_{plan-specific-error}}$$

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*)

Using the parameter estimates from the beta-binomial model, we computed individual plan (or contract) reliability scores. Table 3 shows the distribution of the plan contract-level scores for Medicare, and Table 4 shows the plan-level scores for Medicaid.

Table 3. Plan Contract Reliability Score Distribution for 2015 Medicare Sample

Statistic	Values
Mean	0.7730
Standard Deviation	0.1601
Minimum	0.3628
25th Percentile	0.6569
50th Percentile	0.7995
75th Percentile	0.9153
Maximum	0.9986
Interquartile Range	0.2584

The mean reliability score for the Medicare plan-contracts is 0.7730, and the median is 0.7995. Reliability is affected in part by sample size, and as shown for the Medicare contracts distribution in Table 1, the median plan-contract size is 353 beneficiaries.

In contrast, the median plan distribution for the Medicaid population is much larger – 2,561 beneficiaries (see Table 2). Medicaid plans have very high reliability scores. The mean reliability score in the Medicaid plans is 0.9393, and the median is 0.9926 (see Table 4).

Table 4. Plan Reliability Score Distribution for 2014 Medicaid MAX Sample

Statistic	Values
Mean	0.9393
Standard Deviation	0.1206
Minimum	0.4049
25th Percentile	0.9457
50th Percentile	0.9926
75th Percentile	0.9982
Maximum	1.0000
Interquartile Range	0.0525

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?*)

A reliability score of 0.7 is the minimum threshold for reliability. Based on the mean reliability score of 0.77 for Medicare and 0.94 for Medicaid, the measure is considered reliable.

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

<u>Systematic assessment of face validity</u>

PQA uses a systematic, transparent, consensus-based measure development, testing, and endorsement process. That process used in 2016 to develop this measure is outlined below:

- **Step 1**: Measure concepts for development are prioritized by PQA staff based on input from PQA's Measure Advisement Group, Implementation Advisory Panel, and Patient and Caregiver Advisory Panel. Environmental scans are conducted to identify whether similar measures exist, ensuring harmonization and avoiding duplication. Selected concept ideas are considered to represent areas in which there are measurement and performance gaps to have the greatest chance of implementation in existing measure sets and performance systems, and to align with the National Quality Strategy.
- **Step 2**: PQA Measure Development Teams (MDTs) and Task Forces (TFs), comprised of experts in all phases of drug use and management, discuss and draft specifications for measure concepts that may be appropriate for development into fully specified performance measures. The MDTs/TFs focus on specific aspects of the medication-use system and/or specific therapeutic areas and benefit by having their development work reviewed by larger groups, Stakeholder Advisory Panels. They may also receive

input from the Patient & Caregiver Advisory Panel, Implementation Advisory Panel, and Risk Adjustment Advisory Panel.

- **Step 3**: PQA MDTs/TFs recommend measure concepts to the PQA Quality Metrics Expert Panel (QMEP) for evaluation and refinement. The QMEP reviews the measure concepts to provide an initial assessment of the key properties of performance measures (i.e., importance, scientific acceptability, feasibility and usability). The measure concepts that are rated highly on these key properties will undergo testing and possibly further technical specification as draft measures.
- **Step 4**: The draft measures are provided to PQA member organizations for their comments prior to preparing technical specifications (including National Drug Code [NDC] lists) for pilot testing. PQA staff use member comments and MDT/TF and QMEP recommendations to formulate a testing plan for each draft measure.
- **Step 5**: PQA selects partners to test the draft measures. These partners are often PQA member health plans or academic institutions with expertise in quality and performance measure testing that also have access to the data sources needed to calculate the measure rates. The testing partner implements the draft technical specifications within their existing datasets and provides a report to PQA that details testing results and recommendations for modifications of the technical specifications.
- **Step 6**: The QMEP reviews the testing results and recommendations and determines final criteria for the measure based on the findings. The QMEP provides a final assessment of the feasibility and reliability of the draft measures.
- **Step 7**: The Measure Validity Panel, an independent group of individuals not involved in the development or review of the measure concept or draft measure, determines through discussion and vote whether the performance measure score is an accurate reflection of quality and can distinguish good from poor performance (i.e., face validity).
- **Step 8**: Performance measures that are recommended by the QMEP for endorsement consideration by the PQA membership are posted on the PQA web site for member review, written comments are requested, and a webinar for member organizations is held to gather feedback and address any questions. This process allows members to discuss their views on the measures in advance of the voting period.
- **Step 9**: PQA member organizations vote on endorsement of the performance measures.

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

The measure was assessed for face validity (i.e., whether it appears to measure what it intends to measure) through review by the team that developed the measure (PQA Measure Development Team [MDT] 13: Concurrent Use of Opioids and Benzodiazepines), the PQA Quality Metrics Expert Panel (QMEP), the Measure Validity Panel (MVP), and PQA's full membership. In addition, feedback about validity of the measure was sought out by the two PQA member organizations who tested the measure using their own data, and four external subject matter experts.

MDT 13 was composed of 27 PQA members. After the MDT completed development of the measure specifications, the group voted to determine if the measure concept should continue with further development and review by the PQA QMEP. Out of 27 members of the MDT who voted, 92.5% recommended that the measure move on for QMEP review.

The PQA QMEP is a panel that includes individuals with expertise and experience in pharmacy, medicine, research, and clinical or other technical expertise related to quality improvement and measure development. The names and credentials of the 21 QMEP members in 2016 are listed in Table 5. The QMEP reviewed the measure prior to testing to ensure the importance and usefulness of the draft measure. Specifically, they confirmed that evidence supported that concurrent prescribing of opioids and benzodiazepines was common and associated with overdose deaths. The QMEP reviewed the results of the measure testing including the

performance measure scores reported by plans referenced in Section 2b4 (below). Out of the 16 members of the QMEP who voted, 93.8% recommended that the draft measure be considered for endorsement by the PQA membership, considering the criteria of importance, scientific acceptability, feasibility, and usefulness.

Table 5. PQA 2016 Quality Metrics Expert Panel (QMEP)

QMEP Member Name	QMEP Member Organization
Amanda Brummel, PharmD	Fairview
Bimal Patel, PharmD	MedImpact
Catherine Coast, PharmD	Highmark
Christopher Dezii, RN, MBA, CPHQ	Bristol-Myers Squibb
Christopher Powers, PharmD	CMS
Craig Schilling, PharmD	Optum/UHG
David Nau, PhD, RPh, CPHQ	PQS
Gary Erwin, PharmD	CVS Health
Jenny Weber, PharmD, MS, PCPS, CGP, BCACP	Humana
Jessica Frank, PharmD	OutcomesMTM
Karen Farris, PhD	University of Michigan
Keith Widmer, RPh, BCPP	Express Scripts
Kent Summers, PhD, RPh	Astellas
Lynn Deguzman, PharmD, CGP	Kaiser Permanente
Mary Ann Kliethermes, PharmD	Midwestern University
Mitzi Wasik, PharmD, BCPS	Coventry Health Care/Aetna
Pat Gleason, PharmD, BCPS	Prime Therapeutics
Steve Riddle, PharmD, BCPS	Wolters Kluwer Health
Steven Burch, PhD, RPh	GlaxoSmithKline
Tony Willoughby, PharmD	HealthMart-McKesson
Tripp Logan, PharmD	MedHere Today

After QMEP approval, the draft measure was reviewed by the MVP. The MVP is made up of an independent group of individuals not involved in the development or review of the measure concept or draft measure. Through discussion and vote, the MVP determines whether the performance measure scores have face validity. Of the 6 MVP members who voted, 100% agreed or strongly agreed that the scores obtained from the measure as specified will provide an accurate reflection of quality, and can be used to distinguish good and poor quality between health plans.

PQA membership was notified in November 2016 of the opportunity to consider and vote on endorsement of the performance measure. (Note: PQA membership is comprised of health plans, community pharmacy, long-term care pharmacies, health information technology companies, pharmacy benefit managers, healthcare quality and standards organizations, professional and trade associations, government agencies, and others.) Members received the measure description, key points and supporting evidence, measure specifications, and the performance measure scores reported by the plans. Voting options included, “Agree” (indicating that the

organization approved endorsement of the measure), “Disagree” (indicating that the organization opposed endorsement of the measure) and “Abstain.” Out of the 93 PQA member organizations that cast a vote either in favor of or opposed to endorsement, 89% voted in favor of endorsing the measure.

In addition to this process, 100% of the two PQA member organizations who tested the measure using their own data strongly agreed that the measure reflected the quality of care provided for their population.

The opinion of four subject matter experts was sought in July 2016 for input on the measure elements and assessment of the measure overall. The experts were: Deborah Dowell, MD, MPH, Centers for Disease Control & Prevention; Christopher Jones, PharmD, US Department of Health and Human Services; Joshua Sharenstein, MD, Associate Dean, Johns Hopkins Bloomberg School of Public Health; and Don Teater, MD, Teater Health Solutions (previously, National Safety Council). All four subject matter experts were strongly supportive of the measure.

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

Based upon the systematic, consensus-based PQA measure development process designed to assure face validity, the measure has been determined to have face validity.

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*)

Patients at end of life, undergoing hospice care, and those with cancer may have unusual requirements for pain management. Thus, these are excluded from these measures whenever data are available.
Testing was performed for the hospice exclusion by identifying the number of members in hospice, where available, and determining the percent of the overall population that would be affected by including patients in hospice care.

Cancer exclusions were identified in the Medicare and Medicaid populations using ICD-9 and ICD-10 codes, depending on the time period of the data (ICD-10 coding began in October 2015). Testing involved identifying the number of exclusions, and determining the percent of the overall population that would be affected by including patients with cancer diagnoses.

The exclusions of hospice and cancer are consistent with the 2016 CDC Guideline for Prescribing Opioids for Chronic Pain, which does not apply to active cancer treatment, palliative care, and end-of life treatment because of the unique therapeutic goals, ethical considerations, opportunities for medical supervision, and balance of risks and benefits with opioid therapy in such care.

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*)

For the Medicare population, after applying the age, continuous enrollment and opioid prescription criteria, the hospice patient exclusions ranged from 0.0% to 27.0% among plan contracts, and the cancer exclusions among plan contracts ranged from 0.0% to 52.8%.

For the Medicaid MAX population, after applying the age, continuous enrollment and opioid prescription criteria, the hospice patient exclusions ranged from 0.0% to 1.3% among plans, and the cancer exclusions among plans ranged from 0.0% to 4.5%.

For the one state-based Medicaid program, only one plan was able to identify 3 hospice patients. The cancer exclusion rate was about 4.5% across the three plans.

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*)

The Medicare population shows significant impact of the hospice and cancer exclusions. For the hospice exclusion, up to 27% of beneficiaries in some plan contracts were affected by this exclusion, and the cancer exclusion showed that for some plan contracts, more than half of the population would be affected by this exclusion. Without applying these exclusions, these beneficiaries would be included in the measure. These are significant proportions of the population that could potentially impact the measure rate.

For the Medicaid populations, at the plan level, most of the plans did not identify a substantial number of hospice patients – therefore, no inferences can be drawn from this exclusion. The cancer exclusion had a higher impact. The results show that in some plans, almost 5% of the population has cancer and would be included in the measure if cancer was not excluded. This is a significant proportion of the population that could potentially impact the measure rate.

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** 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)*

To assess significant differences in measure rates, the data described in sections 1.5 and 1.6 above were used to calculate the mean, median, standard deviation, and interquartile range for the measure rates for the Medicare and Medicaid (MAX) populations. In addition, the rates were divided into quartiles, and a Student's t-test was used to compare the rates of the plans in the 25th percentile to the rates of the plans in the 75th percentile.

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)

Tables 6 and 7 show the distribution of the measure rates for the Medicare population. The mean rate was 22.2%, with a median rate of 21.4%, with the lowest plan contract rate at 2.1% and the highest plan contract rate of 44.7%.

Table 6. Variation in Measure Rates – 2015 Medicare Sample

Mean	Median	Standard Deviation
22.2%	21.4%	7.3%

Table 7. Interquartile Range of Measure Rates – 2015 Medicare Sample

Statistic	Value
Minimum	2.1%
25th percentile	17.4%
50th percentile	21.4%
75th percentile	27.3%
Maximum	44.7%
Interquartile Range	9.9%
Student's t-test p-value	<.0001

Tables 8 and 9 show the distribution of the measure rates for the Medicaid MAX population. The mean rate was 5.0%, with a median rate of 4.5%. The lowest plan contract rate was 0.0% and the highest plan contract rate was 17.3%.

Table 8. Variation in Measure Rates – 2014 Medicaid MAX Sample

Mean	Median	Standard Deviation
5.0%	4.5%	3.5%

Table 9. Interquartile Range of Measure Rates – 2014 Medicaid MAX Sample

Statistic	Value
Minimum	0.0%
25th percentile	2.4%
50th percentile	4.5%
75th percentile	6.9%
Maximum	17.3%
Interquartile Range	4.5%
Student's t-test p-value	<.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?*)

For the Medicare population, the measure rates showed significant variation, with a standard deviation of 7.3% and an Interquartile Range of 9.9%. There is a statistically significant difference in measure rates between the top and bottom quartile of the plans included in the testing ($P < .0001$ at alpha = 0.05). This variation shows that there are statistically significant and clinically meaningful differences in rates across plans.

For the Medicaid population, the measure rates showed significant variation, with a standard deviation of 3.5% and an Interquartile Range of 4.5%. There is a statistically significant difference in measure rates between the top and bottom quartile of the plans included in the testing ($P < .0001$ at alpha = 0.05). This variation shows that there are statistically significant and clinically meaningful differences in rates across plans.

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*)

With the use of prescription claims data as the data source for this measure, the dispensing information (including medication, days' supply, quantity dispensed, and dosage) is available for each patient.

Since each of these data elements are available via prescription claims data, it is not expected—nor was it found—that missing data would result. Age is derived from the date of birth in the enrollment data. The date of birth in the CMS Medicare Beneficiaries Summary Files (MBSF) and Medicaid administrative data is considered to largely be valid and reliable since it determines eligibility for enrollment and payment of services.

Patients in hospice are excluded from this measure. No testing was performed on this exclusion as the data source, prescription claims data, do not contain claims for palliative medication, such as opioids, for persons in Medicare Part D that are in hospice care. For the Medicaid population, the majority of the plans were not able to identify hospice exclusions in their data.

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*)

No missing data was found in the testing of this measure.

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*)

As stated above, no missing data was found through testing, nor would missing data be expected to occur in the future. Therefore, performance results would not be biased, as prescription claims data provides the data elements necessary to calculate the measure rate.

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), Other
If other: Medical claims data, Prescription claims data, Prescription Drug Hierarchical Condition Categories (RxHCCs)

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).
N/A

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.
Pilot test sites indicated the measure was feasible and results were able to be reported efficiently, accurately, and without difficulty. The required data (prescription claims and medical claims) are readily available.

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)*.
PQA retains the rights to the measures and can rescind or alter the measures at any time. PQA may approve an organization's use of the measures; however, no organization may use the measures without first obtaining permission from PQA prior to using the measures. Certain uses of the measures are only approved with a licensing agreement from PQA that specifies the terms of use and the licensing fee. PQA reserves the right to determine the conditions under which it will approve use and/or license the measures. Users of the measures shall not have the right to alter, enhance, or otherwise modify the measures.

National Drug Code and ICD code value sets are required to calculate the measure and are provided with the narrative specifications to licensees.

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	Quality Improvement (external benchmarking to organizations) Medicaid Adult Core Measure Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html Quality Improvement (Internal to the specific organization) https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html Medicaid Adult Core Measure Set

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

Program name & sponsor: Centers for Medicare & Medicaid Services (CMS) Medicaid Adult Core Measure Set. National program with state-level voluntary reporting.
Purpose: The Affordable Care Act (Section 1139B) requires the Secretary of Health & Human Services (HHS) to identify and publish a core set of health care quality measures for adult Medicaid enrollees. The core set is published for voluntary use by state Medicaid programs. State data derived from the core measures are part of CMS's annual Child and Adult Core Set measure reporting, which includes publication of datasets that highlight publicly reportable measures. CMS annually releases information on state progress in reporting the Adult Core Set measures and assesses state-specific performance for measures that are reported by at least 25 states and which met internal standards of data quality.
Geographic area: This is a national program with state-level reporting.
Level of measurement and setting: Health plan level of measurement. Outpatient setting.

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?*)
This is a new measure that was developed in 2016. The measure was added to the 2018 Medicaid Adult Core Measure Set; however, measures in the program are publicly reported only if 25 or more states report on the measure. Given that 2018 is the

first year the measure is included, it is not yet publicly reported. We would anticipate adoption of the measure over time, with public reporting once 25 or more states are reporting on the 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 has been added to the Medicaid Adult Core Set for 2018. CMS annually releases information on state progress in reporting the Adult Core Set measures and assesses state-specific performance for measures that are reported by at least 25 states and which met internal standards of data quality.

PQA not only develops and stewards its measures, it also dedicates resources to outreach and implementation efforts. PQA disseminates information regarding the availability of its measures, and provides technical assistance to those implementing or considering implementing PQA-endorsed measures.

Additionally, per the CMS Advance Notice of Methodological Changes for Calendar Year 2019 for Medicare Advantage Capitation Rates, Part C and Part D Payment Policies and 2019 Draft Call Letter (available: https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/Downloads/Advance2019Part2.pdf), CMS proposes to begin reporting the Concurrent Use of Opioids and Benzodiazepines measure in the Medicare Part D Patient Safety reports for the 2018 measurement year, and to add it to the Medicare Part D display page for 2021 (using 2019 data) and 2022 (using 2020 data). CMS also will consider this measure for the 2023 Star Ratings (using 2021 data) pending rulemaking.

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.
N/A

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.**
N/A

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.
N/A

4a2.2.2. **Summarize the feedback obtained from those being measured.**
N/A

4a2.2.3. **Summarize the feedback obtained from other users**
N/A

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.**
N/A

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 performance results can be used to establish benchmarks and identify opportunities to decrease co-prescribing of opioid and benzodiazepines. Sun et al. estimated that the elimination of the concurrent use of opioids and benzodiazepines could reduce the population risk of an emergency department visit or hospital admission for opioid overdose by 15%.(1) Despite the risks, concurrent prescriptions for opioids and benzodiazepines are relatively common and increasing. From 2001-2013, concurrent prescribing increased by nearly 80% (from 9% to 17%) among privately insured patients.(1) In one study, approximately half of the patients received both opioid and benzodiazepine prescriptions from the same prescriber on the same day.(2) In a 2015 analysis of Medicare Part D non-cancer and/or non-hospice patients on opioid therapy, the prevalence of benzodiazepine concurrent use was 24%.(3)

1. Sun EC, Dixit A, Humphreys K, et al. Association between concurrent use of prescription opioids and benzodiazepines and overdose: retrospective analysis. BMJ. 2017;356:j760. doi: 10.1136/bmj.j760. PMID: 28292769
2. Hwang CS, Kang EM, Kornegay CJ, Staffa JA, Jones CM, McAninch JK. Trends in the Concomitant Prescribing of Opioids and Benzodiazepines, 2002-2014. Am J Prev Med. 2016:1-10. doi:10.1016/j.amepre.2016.02.014.
3. CMS. Concurrent Use of Opioids and Benzodiazepines in a Medicare Part D Population. May 12, 2016. 2016. https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Concurrent-Use-of-Opioids-and-Benzodiazepines-in-a-Medicare-Part-D-Population-CY-2015.pdf. Accessed December 6, 2016.

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.

N/A

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

N/A

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)

2940 : Use of Opioids at High Dosage in Persons Without Cancer
2950 : Use of Opioids from Multiple Providers in Persons Without Cancer
2951 : Use of Opioids from Multiple Providers and at High Dosage in Persons Without Cancer

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

Related measures:

- Use of Opioids at High Dosage (NCQA)
- Use of Opioids from Multiple Providers (NCQA)

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 PQA opioid measures (NQF # 2940, 2950, and 2951) use the same target population (denominator), and each have different areas of focus (numerator) related to opioid prescribing. The NCQA opioid measures were developed as an adaptation to existing PQA measures; the NCQA opioid measure denominators are similar to the PQA opioid measures, but have a different area of focus than the concurrent use of opioids and benzodiazepines measure.

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.)
There are no competing measures (i.e., those that addresses both the same measure focus and the 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):** PQA, Inc.
Co.2 **Point of Contact:** Lynn, Pezzullo, LPezzullo@PQAalliance.org, 401-474-9706-
Co.3 **Measure Developer if different from Measure Steward:** PQA, Inc.
Co.4 **Point of Contact:** Lynn, Pezzullo, LPezzullo@PQAalliance.org, 401-474-9706-

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.
A diverse group of stakeholders, including health plans and PBMs (those organizations that will be measured) were well represented throughout the entire development process, including contributing to defining the specifications as members of the Measure Development Team, as testers using the measure specifications to calculate the rates, in the review for face validity and review of testing results as members of the Quality Metrics Expert Panel, and in the vote for PQA endorsement.

PQA Measure Development Teams are small, technically proficient teams composed of diverse stakeholders, to develop individual metrics. Measure Development Team 13 (MDT 13) developed the Concurrent Use of Opioids and Benzodiazepines measure. The members of MDT 13 and their corresponding organizations are listed below:

Cyndi Barham, PharmMD
Maribeth Bettarelli, CVS Health
Donna Boreen, BCBSMN
Jeffrey Bratberg, University of Rhode Island College of Pharmacy (representing the American Association of Colleges of Pharmacy)
Sara Burnheimer, UPMC Health Plan
Pauline Chan, California Department of Health Care Services
Alexandra Cruz, Healthfirst
Samuel Currie, Horizon NJ Health (representing the Association for Community Affiliated Plans)
Tiffany Del Rosario, SCAN Health Plan
Angela DeVeaugh-Geiss, Purdue Pharma LP
Jeff Fink, Express-Scripts
Rainelle Gaddy, Humana
Travis Gau, Medication Management Solutions
Adriane Irwin, American Association of Colleges of Pharmacy
Shellie Keast, University of Oklahoma
Richard Logan, MedHere Today
Michael Long, APhA
Denis Matsuoka, Kaiser Permanente
Karen McLin, SinfoniaRx
Alina Meile, Aetna
Mary Miller, Rite Aid
Anna Polk, Centers for Medicare & Medicaid Services
Madeline Ritchie, Aetna (representing the Academy of Managed Care Pharmacy)
Jennifer Shin, OptumRx
Mindy Smith, PrescribeWellness
Jennifer Snyders, Cigna-HealthSpring
Kathleen Vest, Midwestern University Chicago College of Pharmacy

PQA´s Measure Validity Panel (MVP) is a small group of individuals appointed by PQA staff, to determine whether the performance scores resulting from the measure can be used to distinguish good from poor quality clinical care (i.e., validity). The MVP members that reviewed the Concurrent Use of Opioids and Benzodiazepines measure and their corresponding organizations are listed below:

Susan Skledar, University of Pittsburgh
Ben Banahan, University of Mississippi
Jeff Pohler, University of FL College of Pharmacy
Dan Rehrauer, HealthPartners
Kyle Null, Takeda
Marybeth Farquhar, URAC

PQA´s Quality Metrics Expert Panel (QMEP) is a small group of individuals, selected by PQA staff through an application process, to recommend measure concepts for testing, review measure testing results, and recommend measures for endorsement consideration by PQA membership. The QMEP members that reviewed the Concurrent Use of Opioids and Benzodiazepines measure and their corresponding organizations are listed below:

Amanda Brummel, Fairview Health Services
Bimal Patel, MedImpact
Catherine Coast, Highmark
Christopher Dezii, Bristol-Myers Squibb Company
Christopher Powers, Centers for Medicare & Medicaid Services
Craig Schilling, Optum
David Nau, Pharmacy Quality Solutions
Gary Erwin, Omnicare
Jenny Weber, Humana
Jessica Frank, OutcomesMTM
Karen Farris, University of Michigan College of Pharmacy

Keith Widmer, Express Scripts
Kent Summers, Astellas
Lynn Deguzman, Kaiser Permanente, Northern California
Mary Ann Kliethermes, Midwestern University
Mitzi Wasik, Aetna
Pat Gleason, Prime Therapeutics
Steven Riddle, Wolters Kluwer Health
Steven Burch, GlaxoSmithKline
Tony Willoughby, McKesson
Tripp Logan, Logan and Seiler

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 02, 2018

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 02, 2019

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 3400
Measure Title: Use of pharmacotherapy for opioid use disorder (OUD)
Measure Steward: Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services
Brief Description of Measure: The percentage of Medicaid beneficiaries ages 18 to 64 with an OUD who filled a prescription for or were administered or ordered an FDA-approved medication for the disorder during the measure year. The measure will report any medications used in medication-assisted treatment of opioid dependence and addiction and four separate rates representing the following types of FDA-approved drug products: buprenorphine; oral naltrexone; long-acting, injectable naltrexone; and methadone.
Developer Rationale: Of the 52,404 drug overdose deaths in the United States in 2015, 33,091 (63.1 percent) were due to opioid use (Rudd, Seth, David, & Scholl, 2016) and an estimated 2.5 million individuals have an OUD for abuse or dependence with most not receiving treatment or not receiving the most effective care (Substance Abuse and Mental Health Services Administration, 2015). Among the outcomes that may be affected by OUD treatment are a reduction in drug use, medical problems, and criminal activity and improvements in vocational skills, employment, family relationships, and social activities (Center for Substance Abuse Treatment, 2005). Implementation of new treatment models to expand OUD treatment have been shown to be effective in increasing treatment capacity which is expected to influence patient outcomes (Brooklyn & Sigmon, 2017; Stoller, 2015). It is envisioned that the use of the measure, Use of Pharmacotherapy For Opioid Use Disorder, will improve quality of care by increasing the rate of pharmacotherapy among individuals with an OUD.

There is evidence that pharmacotherapy is related to improved outcomes, therefore, a quality measure to increase access to pharmacotherapy is expected to yield better care for beneficiaries with an OUD. Staying in methadone treatment has been associated with a reduced risk of death (Cousins et al., 2016). Several studies have shown that methadone is safe and effective, especially when higher doses (= 80mg/day) are provided (American Psychiatric Association, 2010). A meta-analysis that reviewed 11 studies on the effectiveness of methadone (Mattick, Breen, Kimber, & Davoli, 2003) found that methadone treatment was more effective than nonpharmacological treatment in retaining clients and reducing their opioid use. Another meta-analysis that reviewed 7 randomized controlled trials and 2 quasi-experimental studies of methadone maintenance found a high level of evidence that methadone treatment had a positive impact on retention in treatment and reduction in opioid use (Fullerton et al., 2014).

Sufficient evidence points to the safety and efficacy of buprenorphine for the treatment of OUD (Parran et al., 2010). The risk of fatal overdose on buprenorphine is substantially lower than that associated with the use of other opioid medications such as methadone because of the ceiling effects of buprenorphine across a wide range of doses (American Society of Addiction Medicine, 2015). One study found that buprenorphine at higher doses (16 to 31mg) is as effective as methadone in reducing opioid use and improve treatment retention (Thomas et al., 2014).

A 2006 Cochrane review and 2009 update found oral naltrexone maintenance therapy alone or associated with psychosocial therapy to be more efficacious than placebo alone or associated with psychosocial therapy in limiting the use of heroin during the treatment, but not in improving retention, or preventing relapse (Minozzi et al., 2006). While oral naltrexone remains an FDA-approved medication for OUD, it has not been widely used due to concerns about adherence(ASAM Practice Guidelines, 2015) and need to maintain withdrawal prior to use (Center for Substance Abuse Treatment, 2005)

In a 6-month multisite double-blind, placebo-controlled RCT conducted in Russia, extended release naltrexone (XR-NTX) was found to be more efficacious than oral NTX with respect to treatment retention and reduction in use of illicit opioids (Krupitsky et al., 2012) and in a 1-year open-label extension of the original trial, about 51% of those who completed the extension were abstinent from opioids at all assessments during the 1-year open-label phase (Krupitsky et al., 2013). XR-NTX was also found to be effective in promoting abstinence across a range of demographic and baseline severity characteristics (Nunes et al., 2015).

Numerator Statement: Beneficiaries ages 18 to 64 with an OUD who filled a prescription for or were administered or ordered an FDA-approved medication for the disorder during the measure year.
Denominator Statement: Number of Medicaid beneficiaries with at least one encounter with a diagnosis of opioid abuse, dependence, or remission (primary or other) at any time during the measurement year.
Denominator Exclusions: None.

Measure Type: Process
Data Source: Claims
Level of Analysis: Population : Regional and State

New Measure - Preliminary Analysis

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 business case logic model outlining that the potential benefits of implementing the measure outweigh the potential costs or unintended consequences of implementing the measure, as well as a clinical practice guideline and 6 systematic reviews:
    - American Society of Addiction Medicine (2015) National practice guideline for the use of medications in the treatment of addiction involving opioid use. **ASAM does not provide grades.**

- Fullerton et al (2014) Medication-Assisted Treatment With Methadone: Assessing the Evidence. **Overall evidence rating for Methadone Maintenance Treatment (MMT) is high.**
- Thomas et al (2014) Medication-Assisted Treatment With Buprenorphine: Assessing the Evidence **Overall evidence rating for Buprenorphine Maintenance Treatment (BMT) is high.**
- Mattick et al (2014) Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence, **High to moderate grades to 31 RCTs under review.**
- Mattick et al (2009) Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence, **High to moderate grades to 11 studies under review.**
- Center for Substance Abuse Treatment (2005) Medication-Assisted Treatment for Opioid Addiction in Opioid Treatment Programs. Treatment Improvement Protocol (TIP) Series 43
- Center for Substance Abuse Treatment (2004) Clinical guidelines for the use of buprenorphine in the treatment of opioid addiction. Treatment Improvement Protocol (TIP) Series 40

Questions for the Committee:

- *What is 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 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

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

- Developer demonstrates performance gap with measure testing results based on 2014 Medicaid Analytic extract data on 16 states. Number of beneficiaries with at least one opioid abuse, dependence, or in remission diagnosis varied across states (1,197 – 59,175).
  - Overall performance rate for any pharmacotherapy use was 57.2%
  - State-level scores ranged from 13.1% - 76.5%

Disparities

- Developer provides pharmacotherapy rates by Medicaid beneficiary eligibility category: age; gender; race/ethnicity; and urban/rural. Results show significant variation in performance rates for each population group tested.

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:

**Evidence is clear. Pharmacotherapy for the treatment of OUD is proven effective.

**The evidence is tangential. There is good evidence that methadone, naltrexone, and buprenorphine are effective in treatment OUD. There is no cited evidence that increasing the prevalence of use among a population improves health outcomes in a linear fashion. The developers might look to the experience in Europe to see if this information exists. Is this measure being used in Dashboards in VT and other states? What is the experience from those states?

1b. Performance Gap

Comments:

**There is a significant performance gap both among states and population groups. Data presented by the developer compelling.

**The SAMHSA publication and table cited to support the treatment gap (Table 7.50A) only shows prevalence rates and does not show use therefore does not support the statement "with most not receiving treatment or not receiving the most effective care" (SAMHSA, 2015). This reference should be deleted. The Max data analyses do show significant variation in use of medications for OUD which does indicate a performance gap, with the caveat that some states may be paying for methadone with state funding and therefore there use will not be apparent in Medicaid claims data. Also, the performance gap analyses included Opioid abuse and opioid dependence in remission which should be excluded and would reduce the performance gap.

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.

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)?*
- *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**

Committee pre-evaluation comments

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

2a1. Reliability – Specifications

Comments:

**The measure is for Medicaid Beneficiaries 18-64 yet Suboxone is also indicated for individuals as young as 16 years old with an OUD. There are no exclusions for individuals with an OUD who do not meet clinical criteria for MAT. The phase methadone "prescription" is misleading--since methadone is often administered by licensed treatment programs. However, the codes listed do include methadone administration. H0020 is not the only code used for methadone administration in a licensed program? Some states use state specific billing codes as the developer mentions. "In State J and State I, 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, State J 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."

2a2. Reliability – Testing

Comments:

**The SUD-4 was highly reliable in terms of ability to distinguish the measure's performance in different states, with an average reliability score of 0.998 across states and a range from 0.993 to 0.999.

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

Comments:

**In State J and State I, the developer 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, State J frequently uses state-specific procedure codes. In the measure submission form, the developer advises 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. Also some states do not cover Methadone under their Medicaid program.

**Yes, missing data on use of methadone paid for by means other than Medicaid is a threat to the validity of the measure. The validity testing examined the correlation between 2 access/use measures. It would have been better and more convincing determination that the measures captures quality to test whether increased prevalence of use of medications resulted in improved health outcomes (ie. reduced overdose deaths).

2b2-3. Other Threats to Validity
2b2. Exclusions
2b3. Risk Adjustment
Comments:
**N/A
**The following diagnoses should be excluded: non-dependent opioid abuse in remission (304.53) and opioid dependence in remission (304.03), opioid abuse (305.5, 305.51, 305.52). Suboxone is only indicated for Opioid Dependence. This specification is encouraging off-label use which has not been approved by the FDA. Also, the inclusion of opioid dependence in remission implies that patients should never tapper off of opioid medications, which is inconsistent with the evidence and clinical recommendations.

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.

- Measure is coded by someone other than person obtaining original information
- All data elements are in defined fields in electronic claims
- There are no fees or licensing requirements to use this measure, which is in the public domain

Questions for the Committee:
- *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:
**This measure requires gathering data from a variety of different data sources and may be complex for certain states to gather.

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**

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

Accountability program details

- CMS is considering implementation plans for this measure. There are no identified barriers to implementation in a public reporting or accountability application.
- The measure is intended for voluntary use by states to monitor and improve the quality of care provided for Medicaid beneficiaries with substance use disorders.

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**:*

- *How 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.

- Adoption of this measure has the potential to improve the quality of care for Medicaid beneficiaries who have an OUD.
    - Overall rate of pharmacotherapy is 57.2% across 16 states included in testing

- 13.05% in State D to 76.59% in State O.

Improvement results

- N/A new measure

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

- Measure has not been implemented yet. No unexpected findings identified during testing.

Potential harms

- Measure has not been implemented yet. No unexpected findings identified during testing.

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**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency

Comments:

**Is this measure being used in VT or other states? Could we learn from their experience?

4b1. Usability – Improvement

Comments:

**The unintended harm is that the medications will be over-prescribed. These medications are have significant risks such as overdose and dependence. There is also increased risk of diversion. Third, there is increased risk of accidental overdose by children. These risks could be partially addressed by removing the population from the denominator for whom the medications are not indicated. Also, by giving guidance on what a reasonable target rate of penetration would be. For example, the Max analyses presented indicate that VT already has a prevalence rate of use of 75%, after 75% more harms may incur. Third, the use of the measure should be accompanied by surveillance to detect any unintended harms such as increased diversion, overdose and the potential for risk in using the measure should be acknowledged. When the measure comes up for maintenance, data on unintended consequence should be presented.

Criterion 5: Related and Competing Measures

Related or competing measures

- 3175 : Continuity of Pharmacotherapy for Opioid Use Disorder
- Evidence of medication-assisted treatment (MAT) among patients with opioid use disorder (OUD) or OD, Steward: OptumLabs

Harmonization

- Developer states measure specifications have been harmonized to the extent possible with above related measures.

Public and member comments

Comments and Member Support/Non-Support Submitted as of: June 7, 2018

- No comments received.
- No NQF Members have submitted support/non-support choices as of this date.

Measure Number: 3400

Measure Title: Use of pharmacotherapy for opioid use disorder (OUD)

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

Instructions for filling out this form:

- Please complete this form for each measure you are evaluating.
- Please pay close attention to the skip logic directions. ***Directives that require you to skip questions are marked in red font.***
- If you are unable to check a box, please highlight or shade the box for your response.
- You must answer the "overall rating" item for both Reliability and Validity. Also, be sure to answer the composite measure question at the end of the form if your measure is a composite.
- For several questions, we have noted which sections of the submission documents you should ***REFERENCE*** and provided ***TIPS*** to help you answer them.
- ***It is critical that you explain your thinking/rationale if you check boxes that require an explanation.*** Please add your explanation directly below the checkbox in a different font color**.** Also, feel free to add additional explanation, even if you select a checkbox where an explanation is not requested (if you do so, please type this text directly below the appropriate checkbox).
- ***Please refer to the Measure Evaluation Criteria and Guidance document (pages 18-24) and the 2-page Key Points document when evaluating your measures***. This evaluation form is an adaptation of Alogorithms 2 and 3, which provide guidance on rating the Reliability and Validity subcriteria.
- ***Remember*** that testing at either the data element level **OR** the measure score level is accepted for some types of measures, but not all (e.g., instrument-based measures, composite measures), and therefore, the embedded rating instructions may not be appropriate for all measures.
- ***Please base your evaluations solely on the submission materials provided by developers.*** NQF strongly discourages the use of outside articles or other resources, even if they are cited in the submission materials. If you require further information or clarification to conduct your evaluation, please communicate with NQF staff (methodspanel@qualityforum.org).

RELIABILITY

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

   REFERENCE: "MIF_xxxx" document
   ***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?

   REFERENCE: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2
   ***TIPS**: Check the "NO" box below if: only descriptive statistics are provided; only describes process for data management/cleaning/computer programming; testing does not match measure specifications (i.e., data source, level of analysis, included patients, etc.)*

   ☒Yes (go to Question #3)

   ☐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, skip Questions #3-8, then go to Question #9)

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

   REFERENCE: "Testing attachment_xxx", section 2a2.1 and 2a2.2
   ***TIPS**: Answer no if: only one overall score for all patients in sample used for testing patient-level data*

   ☒Yes (go to Question #4)

   ☐No (skip Questions #4-5 and go to Question #6)

4. 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.*

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #5)

   ☐No (please explain below, then go to question #5 and rate as INSUFFICIENT)
   Signal-to-noise ratio used to assess variation between state scores and temporal consistency assessed with Spearman rank correlation.

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

   REFERENCE: Testing attachment, section 2a2.2
   ***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 #6)

☐Moderate (go to Question #6)
☐Low (please explain below then go to Question #6)
☐Insufficient (go to Question #6)

Average reliability score of 0.998 across states and a range from 0.993 to 0.999. Spearman rank correlation of state-level measure rate between CY 2013 and 2014 is 0.92 at the 95 percent confidence interval (0.77, 0.97).

6. Was reliability testing conducted with patient-level data elements that are used to construct the performance measure?
   REFERENCE: Testing attachment, section 2a2.
   ***TIPS**: Prior reliability studies of the same data elements may be submitted; if comparing abstraction to "authoritative source/gold standard" go to Question #9)*
   ☐Yes (go to Question #7)
   ☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #5, skip questions #7-9, then go to Question #10 (OVERALL RELIABILITY); otherwise, skip questions #7-8 and go to Question #9)

7. Was the method described and appropriate for assessing the reliability of ALL critical data elements?
   REFERENCE: Testing attachment, section 2a2.2
   ***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 #8)
   ☐No (if no, please explain below, then go to Question #8 and rate as INSUFFICIENT)

8. **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?
   REFERENCE: Testing attachment, section 2a2
   ***TIPS**: Consider the following: Is the test sample adequate to generalize for widespread implementation? Can data elements be collected consistently?*
   ☐Moderate (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 as MODERATE)
   ☐Low (skip Question #9 and go to Question #10 (OVERALL RELIABILITY); if score-level testing was NOT conducted, rate Question #10 (OVERALL RELIABILITY) as LOW)
   ☐Insufficient (go to Question #9)

9. Was **empirical VALIDITY testing** of patient-level data conducted?
   REFERENCE: testing attachment section 2b1.
   NOTE: Skip this question if empirical reliability testing was conducted and you have rated Question #5 and/or #8 as anything other than INSUFFICIENT)

TIP: *You should answer this question ONLY if score-level or data element reliability testing was NOT conducted or if the methods used were NOT appropriate. For most measures, NQF will accept data element validity testing in lieu of reliability testing—but* ***check with NQF staff before proceeding, to verify.***

☐Yes (go to Question #10 and answer using your rating from data element validity testing – Question #23)

☐No (please explain below, go to Question #10 (OVERALL RELIABILITY) and rate it as INSUFFICIENT. Then go to Question #11.)

OVERALL RELIABILITY RATING

10. **OVERALL RATING OF RELIABILITY** taking into account precision of specifications (see Question #1) 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, but check with NQF staff]

VALIDITY

Assessment of Threats to Validity

11. Were potential threats to validity that are relevant to the measure empirically assessed ()?

REFERENCE: Testing attachment, section 2b2-2b6

TIPS*: Threats to validity that should be assessed include: exclusions; need for risk adjustment; ability to identify statistically significant and meaningful differences; multiple sets of specifications; missing data/nonresponse.*

☒Yes (go to Question #12)

☐No (please explain below and then go to Question #12) [NOTE that ***non-assessment of applicable threats should result in an overall INSUFFICENT rating for validity***]

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

REFERENCE: Testing attachment, section 2b2.

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 #13)

☐No (go to Question #13)

☒Not applicable (i.e., there are no exclusions specified for the measure; go to Question #13)

13. Analysis of potential threats to validity: Risk-adjustment (this applies to all outcome, cost, and resource use measures and "NOT APPLICABLE" is not an option for those measures; the risk-adjustment questions (13a-13c, below) also may apply to other types of measures)

REFERENCE: Testing attachment, section 2b3.

13a. Is a conceptual rationale for social risk factors included? ☐Yes ☐No

13b. Are social risk factors included in risk model? ☐Yes ☐No

13c. Any concerns regarding the risk-adjustment approach?

***TIPS**: Consider the following: **If measure is risk adjusted**: If the developer asserts there is **no conceptual basis** for adjusting this measure for social risk factors, do you agree with the rationale? 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? If a measure is NOT risk-adjusted, is a justification for **not risk adjusting** provided (conceptual and/or empirical)? Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

☐Yes (please explain below then go to Question #14)

☐No (go to Question #14)

☒Not applicable (e.g., this is a structure or process measure that is not risk-adjusted; go to Question #14)

N/A Process measure

14. Analysis of potential threats to validity: Any concerns regarding ability to identify meaningful differences in performance or overall poor performance?
REFERENCE: Testing attachment, section 2b4.

☐Yes (please explain below then go to Question #15)

☒No (go to Question #15)

15. Analysis of potential threats to validity: Any concerns regarding comparability of results if multiple data sources or methods are specified?
REFERENCE: Testing attachment, section 2b5.

☐Yes (please explain below then go to Question #16)

☒No (go to Question #16)

☐Not applicable (go to Question #16)

16. Analysis of potential threats to validity: Any concerns regarding missing data?
REFERENCE: Testing attachment, section 2b6.

☐Yes (please explain below then go to Question #17)

☒No (go to Question #17)

Assessment of Measure Testing

17. Was empirical validity testing conducted using the measure as specified and with appropriate statistical tests?
REFERENCE: Testing attachment, section 2b1.

***TIPS**: Answer no if: only face validity testing was performed; 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 #18)

☐No (please explain below, then skip Questions #18-23 and go to Question #24)

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

REFERENCE: Testing attachment, section 2b1.

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

☒Yes (go to Question #19)

☐No (please explain below, then skip questions #19-20 and go to Question #21)

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

REFERENCE: Testing attachment, section 2b1.

***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 #20)

☐No (please explain below, then go to Question #20 and rate as INSUFFICIENT)

Convergent validity assessed with Spearman rank correlation using two HEDIS measures.

20. **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 #21)

☐Moderate (go to Question #21)

☐Low (please explain below then go to Question #21)

☐Insufficient (go to Question #21)

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

REFERENCE: Testing attachment, section 2b1.

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

☐Yes (go to Question #22)

☒No (if there is score-level testing that you rated something other than INSUFFICIENT in Question #20, skip questions #22-25, and go to Question #26 (OVERALL VALIDITY); otherwise, skip questions #22-23 and go to Question #24)

22. 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.*

***REFERENCE:** Testing attachment, section 2b1.*

***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 #23)

☐No (please explain below, then go to Question #23 and rate as INSUFFICIENT)

23. **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 (skip Questions #24-25 and go to Question #26)

☐Low (please explain below, skip Questions #24-25 and go to Question #26)

☐Insufficient (go to Question #24 only if no other empirical validation was conducted OR if the measure has not been previously endorsed; otherwise, skip Questions #24-25 and go to Question #26)

24. 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?

NOTE: If appropriate empirical testing has been conducted, then evaluation of face validity is not necessary; you should skip this question and Question 25, and answer Question #26 based on your answers to Questions #20 and/or #23]

REFERENCE: Testing attachment, section 2b1.

***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 #25)

☐No (please explain below, skip question #25, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT)

25. **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?

REFERENCE: Testing attachment, section 2b1.

***TIPS**: Face validity is no longer accepted for maintenance measures unless there is justification for why empirical validation is not possible and you agree with that justification.*

☐Yes (if a NEW measure, go to Question #26 (OVERALL VALIDITY) and rate as MODERATE)

☐ Yes (if a MAINTENANCE measure, do you agree with the justification for not conducting empirical testing? If no, go to Question #26 (OVERALL VALIDITY) and rate as INSUFFICIENT; otherwise, rate Question #26 as MODERATE)

☐No (please explain below, go to Question #26 (OVERALL VALIDITY) and rate AS LOW)

OVERALL VALIDITY RATING

26. **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 instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT—please check with NQF staff if you have questions.]

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Use of pharmacotherapy for opioid use disorder (OUD)
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A
Date of Submission: 3/13/2018

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: Use of pharmacotherapy for opioid use disorder (OUD)

☐ 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.

The logic model suggests that the potential benefits of implementing SUD-4 outweigh the potential costs or unintended consequences of implementing the measure. Benefits to Medicaid beneficiaries include reduction in opioid use, relapses, emergency department or inpatient admissions related to overdoses, and lower mortality (Clark et al., 2015; Clark et al., 2014; Fullerton et al., 2014; R.P. Mattick, Breen, Kimber, & Davoli, 2009; R.P. Mattick, Breen, Kimber, & Davoli, 2014; Parran et al., 2010; M. Pierce et al., 2016; Thomas et al., 2014) . The benefits to society include reductions in costs related to criminal activity and in overall health care costs (Ball & Ross, 1991; Clark et al., 2015). Furthermore, generally, pharmacotherapy has been found to result in lower total health care expenditures (Mohlman, Tanzman, Finison, Pinette, & Jones, 2016).

Appendix A: Use of pharmacotherapy for OUD—Business case logic model

Measure information

Measure description: The percentage of Medicaid beneficiaries with an OUD who filled a prescription for, or were administered, an FDA-approved medication for the disorder. This is a process measure that can be calculated with administrative (eligibility, claim, and pharmacy) data.

Numerator: Number of Medicaid beneficiaries age 18 to 64 with at least one prescription filled, or who were administered one of three FDA-approved OUD treatment medications (buprenorphine, naltrexone, methadone) at any point during the measurement year, identified through pharmacy claims (relevant NDC code) or through relevant HCPCS coding of medical service

Denominator: Number of Medicaid beneficiaries with at least one encounter with a diagnosis of opioid abuse or dependence (primary or other) at any time during the measurement year

Exclusions: Individuals whose only OUD-related diagnosis is "in remission"

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

- Studies show that the percentage of OUD patients who filled or received a prescription for an OUD medication is well below 50%, ranging from 5% to 34%; therefore, there is much room for improvement in the utilization of pharmacotherapy.

Health outcomes (per studies cited in text):

Impact on clients:

- Reduction in opioid use
- Decrease in emergency department or inpatient admissions related to substance use or overdose

Impact on society

- Reduction in crime related to substance use

Health care costs

- Lower health care costs

Costs

Implementation costs

- Low cost to adopt measure since using 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

Intervention costs

- Increased cost to Medicaid if more individuals receive pharmacotherapy treatment
- Cost to client for additional treatment, lost days of work to go to treatment, transportation costs, and possibly child care costs
- Cost for programs to hire additional prescribing practitioners who have a waiver to prescribe

Influencing factors

Factors that influence the adoption, implementation, or endorsement of the measure

- Patient/provider education
- Effort involved with using the measure
- Champion supportive of the measure and quality improvement
- Data availability and completeness

Factors that influence outcomes through use of the measure

- Workforce limits
- State pharmacotherapy policies
- Variations in state coverage of services
- Provider attitudes/concerns

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

****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 1: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Source** • **Title:** National practice guideline for the use of medications in the treatment of addiction involving opioid use • **Author:** American Society of Addiction Medicine • **Date:** Adopted by the ASAM Board of Directors June 1, 2015 • **Citation:** American Society of Addiction Medicine. (2015). National practice guideline for the use of medications in the treatment of addiction involving opioid use. Pages: 4, 24, 29, 32, 36 & 37. • **URL:** https://www.asam.org/docs/default-source/practice-support/guidelines-and-consensus-docs/asam-national-practice-guideline-supplement.pdf?sfvrsn=24#search="national practice guideline supplement"

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.	p. 4, "This Practice Guideline is primarily intended for clinicians involved in evaluating patients and providing authorization for pharmacological treatments at any level." p.24, "Methadone is frequently used to manage withdrawal symptoms from opioids and is recommended for pharmacological treatment of opioid use disorder." p.29, "Treatment with methadone at an OTP is recommended for patients who have opioid use disorder, are able to give informed consent, and have no specific contraindications for agonist treatment." p. 32, "Buprenorphine is recommended for the treatment of opioid use disorder… For this Practice Guideline, recommendations using the term 'buprenorphine' will refer generally to both the buprenorphine only and the combination buprenorphine/naloxone formulations… Buprenorphine is an effective treatment recommended for patients who have opioid use disorder, are able to give informed consent, and have no specific contraindications for agonist treatment." p. 36, "Naltrexone is recommended for pharmacological treatment of opioid use disorder." p. 37, "Naltrexone is a recommended treatment in preventing relapse in opioid use disorder. Oral formula naltrexone may be considered for patients in whom adherence can be supervised or enforced. Extended-release injectable naltrexone may be more suitable for patients who have issues with adherence."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Not Applicable. The guidelines cited above do not provide grades (e.g., USPSTF grades A, B, etc.).
Provide all other grades and definitions from the evidence grading system	The ASAM practice guidelines were developed by using the RAND appropriateness method. Guideline developers reviewed existing literature and guidelines, treatment scenarios, appropriateness ratings, and other documents. An independent panel convened by ASAM oversaw development of the guidelines.
Grade assigned to the **recommendation** with definition of the grade	Not Applicable. The guidelines cited above do not provide grades (e.g., USPSTF grades A, B, etc.).
Provide all other grades and definitions from the recommendation grading system	The ASAM practice guidelines were developed by using the RAND appropriateness method. Guideline

	developers reviewed existing literature and guidelines, treatment scenarios, appropriateness ratings, and other documents. An independent panel convened by ASAM oversaw development of the guidelines.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The number of studies referenced in support of the guidelines was 173. The quality of the evidence selected indicates that included studies were considered to be of high quality by an expert panel.
Estimates of benefit and consistency across studies	Buprenorphine pharmacotherapy is recommended for the treatment of opioid use disorder (American Society of Addiction Medicine, 2015). Buprenorphine relieves drug cravings without producing the euphoria or dangerous side effects of other opioids and is more effective than a placebo (Johnson et al., 1995; Ling et al., 1998). With passage of the Drug Addiction Treatment Act of 2000 (Substance Abuse and Mental Health Services Administration, 2015) and FDA approval of buprenorphine in 2002, buprenorphine became the first pharmacotherapy medication eligible to be prescribed by certified clinicians in an office-based setting. Clinicians may apply for Drug Enforcement Administration (DEA) waivers to prescribe certain narcotic medications, including buprenorphine, expanding the accessibility of community-based treatment options. Methadone at moderate and high doses is effective in reducing illicit opioid use, although higher doses have been found to produce a significantly greater decrease in opioid use (Strain, Bigelow, Liebson, & Stitzer, 1999). Oral naltrexone treatment is recommended for candidates who can be closely supervised and who are highly motivated. Extended-release injectable naltrexone has also been found to be an efficacious treatment for opioid use disorder. It may be especially useful for patients who have contraindications to, or who failed pharmacotherapy with, buprenorphine and methadone.
What harms were identified?	Sufficient evidence points to the safety and efficacy of buprenorphine for the treatment of OUD (Parran et al., 2010). The risk of fatal overdose on buprenorphine is substantially lower than that associated with the use of other opioid medications such as methadone because of the ceiling effects of buprenorphine across a wide range of doses (American Society of Addiction Medicine, 2015).

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	The following more recent articles detail supporting evidence: Pierce, M., Bird, S. M., Hickman, M., Marsden, J., Dunn, G., Jones, A., & Millar, T. (2016). Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. *Addiction*, *111*(2), 298-308. This study compared "the change in illicit opioid users' risk of fatal drug-related poisoning (DRP) associated with opioid agonist pharmacotherapy (OAP) and psychological support". Findings indicated that "patients who received only psychological support for opioid dependence in England appear to be at greater risk of fatal opioid poisoning than those who received opioid agonist pharmacotherapy" (Matthias Pierce et al., 2016). Ayanga, D., Shorter, D., & Kosten, T. R. (2016). Update on pharmacotherapy for treatment of opioid use disorder. *Expert opinion on pharmacotherapy*, *17*(17), 2307-2318. This article reviewed "pharmacologic strategies for OUD treatment, discussing both primary as well as adjunctive therapy modalities." Results indicate that "medication therapy for treatment of OUD has demonstrated efficacy and is of great clinical benefit. While agonist treatment with methadone or buprenorphine remains the gold standard, there is an important place for use of long-acting antagonist therapy with naltrexone" (Ayanga, Shorter, & Kosten, 2016).

Source of Systematic Review 2: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Source** • **Title:** Medication-Assisted Treatment With Methadone: Assessing the Evidence • **Author:** Fullerton, C. A., Kim, M., Thomas, C. P., Lyman, D. R., Montejano, L. B., Dougherty, R. H., Daniels, A. S., Ghose, S. S., & Delphin-Rittmon, M. E. • **Date:** 2014 • **Citation:** Fullerton, C. A., Kim, M., Thomas, C. P., Lyman, D. R., Montejano, L. B., Dougherty, R. H., ... & Delphin-Rittmon, M. E. (2014). Medication-assisted treatment with

	methadone: assessing the evidence. *Psychiatric Services*, *65*(2), 146-157. **URL:** N/A
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.	"Overall, there is a high level of evidence for the effectiveness of [Methadone Maintenance Treatment [MMT)] in improving treatment retention and decreasing illicit opioid use (see box on previous page). Research findings regarding the impact of MMT on many secondary outcomes, such as mortality, drug-related HIV risk behaviors, and criminal activity, are less conclusive but suggest positive trends. Finally, research has not conclusively shown positive impacts on sex-related HIV risk behaviors, nonopioid illicit drug or alcohol use, or other social consequences. Methadone maintenance doses above 60 mg confer greater efficacy in retention and suppression of illicit opioid use; however, there is limited evidence that doses above 100 mg provide additional benefits. No evidence has emerged to delineate the duration of MMT beyond an indefinite period. Although MMT generally is believed to reduce mortality risk among individuals with opioid dependence, methadone is also associated with significant adverse events, such as respiratory depression and cardiac arrhythmias, in the presence of rapid titrations or other risk factors. There is no clear evidence that structured psychotherapy provided in addition to the psychosocial support normally offered at methadone treatment centers conveys additional benefit. MMT improves pregnancy-related outcomes by reducing illicit drug use and increasing treatment retention. However, newborn infants of mothers treated with methadone during pregnancy may be born with NAS irrespective of the methadone dose used by the mothers." "MMT is an important treatment option for opioid dependence. Providers, consumers, and family members should be educated about the benefits of MMT in helping individuals manage opioid use disorders and about appropriate ways to avoid the significant adverse events that can occur with methadone. Providers and consumers need to be educated regarding appropriate doses to improve efficacy and appropriate initiation to minimize adverse events. Because of MMT's relative efficacy, efforts should be made to increase access to MMT for all individuals who struggle with opioid use disorders. Directors of state mental health and substance abuse

	agencies and community health organizations should look for methods to increase access to MMT, and purchasers of health care services should cover appropriately monitored MMT."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	"Because of the large number of trials included as individual studies or as part of review articles, the overall evidence rating for MMT is high. Several meta-analyses, reviews, and RCTs representing more than three independent RCTs have reported on the primary outcomes of MMT, which are retention in treatment and reduction of illicit opioid use. In addition, meta-analyses, reviews, RCTs, and quasi-experimental studies representing more than three RCTs or two RCTs and two quasi-experimental studies have addressed secondary outcomes such as other illicit drug use, HIV risk behaviors, criminal behaviors, heroin craving, and mortality." "Evidence for the effectiveness of methadone maintenance treatment: high. Evidence clearly shows that MMT has a positive impact on: retention in treatment, illicit opioid use. Evidence is less clear but suggestive that MMT has a positive impact on: mortality, illicit drug use (nonopioid), drug-related HIV risk behaviors, criminal activity, evidence suggests that MMT has little impact on: sex-related HIV risk behaviors."
Provide all other grades and definitions from the evidence grading system	"Three levels of evidence (high, moderate, and low) were used to indicate the overall research quality of the collection of studies. Ratings were based on predefined benchmarks that considered the number of studies and their methodological quality. If ratings were dissimilar, a consensus opinion was reached." "High ratings indicate confidence in the reported outcomes and are based on three or more RCTs with adequate designs or two RCTs plus two quasiexperimental studies with adequate designs. Moderate ratings indicate that there is some adequate research to judge the service, although it is possible that future research could influence reported results. Moderate ratings are based on the following three options: two or more quasiexperimental studies with adequate design; one quasi-experimental study plus one RCT with adequate design; or at least two RCTs with some methodological weaknesses or at least three quasi-experimental studies with some methodological weaknesses. Low ratings indicate that

	research for this service is not adequate to draw evidence based conclusions. Low ratings indicate that studies have nonexperimental designs, there are no RCTs, or there is no more than one adequately designed quasi-experimental study."
Grade assigned to the **recommendation** with definition of the grade	See grade of evidence
Provide all other grades and definitions from the recommendation grading system	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	"Authors reviewed meta-analyses, systematic reviews, and individual studies of MMT from 1995 through 2012. Databases searched were PubMed, PsycINFO, Applied Social Sciences Index and Abstracts, Sociological Abstracts, Social Services Abstracts, and Published International Literature on Traumatic Stress." "The literature search found 7 RCTs and two retrospective, quasi-experimental studies. 15 reviews or meta-analyses that examined multiple studies [were also included]."
Estimates of benefit and consistency across studies	"Research supports MMT's positive impact on treatment retention and suppression of heroin use, particularly at higher methadone doses. Findings regarding secondary outcomes are mixed, although there is general support that MMT has a positive impact on criminal activity associated with heroin use, as well as on mortality and risk behaviors for HIV and hepatitis C infection." "In general, these and later studies found that when MMT is provided at adequate dose levels, it is more effective than no medication treatment in retaining patients in treatment and reducing illicit opioid use." "MMT during pregnancy was associated with decreased illicit opioid use, increased rates of prenatal retention in treatment, decreased pregnancy complications, and generally improved fetal outcomes."
What harms were identified?	"MMT has been found to put newborn infants at risk for neonatal abstinence syndrome (NAS)—a condition characterized by dysfunction of the autonomic nervous system, gastrointestinal tract, and respiratory system and by irritability of the central nervous system. NAS often requires detoxification treatment in the hospital with a morphine taper. Reported rates of

	withdrawal symptoms among neonates born to opioid-addicted mothers who continued to use opiates within a week of giving birth range from 55% to 94%, and rates of NAS that develop among neonates as a result of treating the mother with MMT during pregnancy fall into this range. Recent studies on the long-term impact of NAS on development are scant. Older studies indicated no differences in cognitive performance among four-year old children of mothers receiving MMT and children of mothers with similar demographic characteristics in a control group. However, scores of children in both groups were lower than population norms." "Between 1999 and 2004, deaths attributed to methadone increased by 390%. Evidence suggests that this change was largely related to the increased use of methadone for pain analgesia rather than MMT. Nonetheless, the sharp rise of methadone-related deaths highlights safety issues—in particular, the risks of respiratory depression and cardiac QT interval prolongation. The QT interval is a measure of time between the start of the Q wave and the end of the T wave in the heart's electrical cycle that is measured by an electrocardiogram. Prolongation of the QT interval can lead to serious heart arrhythmias such as Torsades de Pointes (TdP) and sudden death." "As a result of this rise in mortality, the U.S. Food and Drug Administration issued a physician safety alert in 2006 highlighting fatalities and cardiac arrhythmias associated with methadone. Respiratory depression is most often a consequence of methadone accumulation and use of concurrent illicit drugs or medications that also suppress the central nervous system. Reviews suggest that initiation into methadone treatment is a particularly vulnerable time in both methadone maintenance and pain therapy populations, particularly if the dose is increased rapidly. The most common drugs associated with respiratory suppression are benzodiazepines and alcohol. Deaths from respiratory depression may also be caused by inappropriate dosing by methadone recipients and by diversion of methadone, which occurs when individuals who have a prescription for methadone sell or give their methadone to others rather than using it themselves."

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See systematic review 1 which supports these recommendations.

Source of Systematic Review 3: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Source** • **Title:** Medication-Assisted Treatment With Buprenorphine: Assessing the Evidence • **Author:** Thomas, C. P., Fullerton, C. A., Kim, M., Montejano, L., Lyman, D. R., Dougherty, R. H., Daniels, A., S., Ghose, S. S., Delphin-Rittmon, M. E. • **Date:** 2014 • **Citation:** Thomas, C. P., Fullerton, C. A., Kim, M., Montejano, L., Lyman, D. R., Dougherty, R. H., ... & Delphin-Rittmon, M. E. (2014). Medication-assisted treatment with buprenorphine: assessing the evidence. *Psychiatric Services*, *65*(2), 158-170. **URL:** N/A
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.	"[Buprenorphine Maintenance Treatment] (BMT)" is associated with improved outcomes compared with placebo for individuals and pregnant women with opioid use disorders." "Policy makers have reason to promote access to BMT for patients in substance use treatment who may wish to choose BMT as a potentially safer alternative to [Methadone Maintenance Treatment (MMT)]. Administrators of substance use treatment programs, community health centers, and managed care organizations and other purchasers of health care services, such as Medicare, Medicaid, and commercial insurance carriers, should give careful consideration to BMT as a covered benefit."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The research designs of the identified studies were examined. Three levels of evidence (high, moderate, and low) were used to indicate the overall research quality of the collection of studies. Ratings were based on predefined benchmarks that considered the number of studies and their methodological quality. If ratings were dissimilar (occurring for 13% of the studies rated), a consensus opinion was reached." "In general, high ratings indicate confidence in the reported outcomes and are based on three or more RCTs with adequate designs or two RCTs plus two quasi-experimental studies with adequate designs.

	Moderate ratings indicate that there is some adequate research to judge the service, although it is possible that future research could influence reported results. Moderate ratings are based on the following three options: two or more quasiexperimental studies with adequate design; one quasi-experimental study plus one RCT with adequate design; or at least two RCTs with some methodological weaknesses or at least three quasi-experimental studies with some methodological weaknesses. Low ratings indicate that research for this service is not adequate to draw evidence based conclusions. Low ratings indicate that studies have nonexperimental designs, there are no RCTs, or there is no more than one adequately designed quasi-experimental study." The grade assigned to the evidence was "high". The author's stated that "because of the large number of trials, the overall evidence for BMT was rated as high. Thus the level of research evidence is similar for BMT and MMT. In addition, multiple meta-analyses, reviews, and more than three independent RCTs have compared BMT with MMT on the primary outcomes stated above, and these results are also based on a high level of evidence in RCTs or reviews. Secondary outcomes, such as use of other illicit drugs, criminal behaviors, and other measures of addiction severity or psychosocial functioning varied among studies; as a result, the evidence for these secondary outcomes is not as strong."
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	"Evidence for the effectiveness of BMT: high. Evidence clearly shows that BMT has a positive impact compared with placebo on: retention in treatment, and illicit opioid use. Evidence is mixed for its impact on: nonopioid illicit drug use." "In general, high ratings indicate confidence in the reported outcomes and are based on three or more RCTs with adequate designs or two RCTs plus two quasi-experimental studies with adequate designs. Moderate ratings indicate that there is some adequate research to judge the service, although it is possible that future research could influence reported results. Moderate ratings are based on the following three options: two or more quasiexperimental studies with adequate design; one quasi-experimental study plus

	one RCT with adequate design; or at least two RCTs with some methodological weaknesses or at least three quasi-experimental studies with some methodological weaknesses. Low ratings indicate that research for this service is not adequate to draw evidence based conclusions. Low ratings indicate that studies have nonexperimental designs, there are no RCTs, or there is no more than one adequately designed quasi-experimental study."
Provide all other grades and definitions from the recommendation grading system	N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	"The literature search revealed 16 RCTs, a randomized cross-over study, a study using a self-administered survey, and a retrospective descriptive study. RCTs used either buprenorphine alone or buprenorphine-naloxone. The search also found seven reviews or meta-analyses."
Estimates of benefit and consistency across studies	"Buprenorphine has a better safety profile than methadone, and the ability to prescribe buprenorphine in office facilities as opposed to only in opioid treatment programs improves access to care and earlier initiation of treatment. A key advantage of buprenorphine is its availability." "Both BMT and MMT improve pregnancy-related outcomes by reducing illicit drug use during pregnancy."
What harms were identified?	"The pharmacology of buprenorphine affords it a better safety profile than methadone, which is important considering that methadone is associated with one-third of opioid-related overdose deaths annually. Because it is a partial agonist at the mu opiate receptor, it has a ceiling effect that limits its potential to cause respiratory depression compared with methadone. However, this risk still exists, especially if buprenorphine is used in combination with other central nervous system depressants such as benzodiazepines or alcohol or is used in higher doses. In addition, unlike methadone, buprenorphine at standard doses does not affect cardiac electrophysiology by lengthening the cardiac QT interval—a mechanism that can lead to serious cardiac arrhythmias. Buprenorphine also has fewer drug interactions than methadone, especially with HIV medications. Taken together, the articles reviewed suggest that the efficacy of BMT is dose dependent, and dose is important to take into account when comparing medications."

	"Infants of mothers treated with buprenorphine during pregnancy may be born with NAS, although NAS appears to be less severe in infants of mothers treated with buprenorphine than of those treated with methadone." "Buprenorphine naloxone retains some potential for abuse intravenously, but the combination has less abuse potential as measured by self-administration than buprenorphine alone or heroin."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See systematic review 1 which supports these recommendations.

Source of Systematic Review 4: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Source** • **Title:** Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. • **Author:** Mattick RP, Breen C, Kimber J, and Davoli M. • **Date:** 2014 • **Citation:** Mattick RP, Breen C, Kimber J, and Davoli M. Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. Cochrane Database of Systematic Reviews 2014, Issue 2. Art. No.: CD002207. DOI: 10.1002/14651858.CD002207.pub4. • **URL:** http://onlinelibrary.wiley.com/doi/-10.1002/14651858.CD002207.pub4/epdf
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 meta-analysis of 31 trials found that the quality of evidence for buprenorphine maintenance varied from high to moderate (R.P. Mattick et al., 2014). The analysis examined randomized controlled trials of buprenorphine maintenance treatment versus placebo or methadone treatment for management of opioid use disorders. Strong evidence indicated that buprenorphine is superior to placebo medication in retention of participants in treatment at all dosing levels considered in the analyses. Specifically, buprenorphine retained participants better than a placebo at low doses (2 to 6 mg), at medium doses (7 to 15 mg), and at high doses (≥ 16 mg). The authors based their conclusion on placebo-controlled trials, concluding that buprenorphine is an effective medication for retaining individuals with an OUD in

	treatment at any dose above 2 mg and for suppressing illicit opioid use (at doses 16 mg or greater).
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The Cochrane review meta-analyses include grades. The meta-analysis of buprenorphine treatment (R.P. Mattick et al., 2014) assigned high to moderate grades to the studies under review.
Provide all other grades and definitions from the evidence grading system	High quality: Further research is very unlikely to change the authors' confidence in the estimated effect. Moderate quality: Further research is likely to have an important impact on confidence in the estimate of effect and may change the estimate.
Grade assigned to the **recommendation** with definition of the grade	The Cochrane review meta-analyses include grades. The meta-analysis of buprenorphine treatment (R.P. Mattick et al., 2014) assigned high to moderate grades to the studies under review.
Provide all other grades and definitions from the recommendation grading system	High quality: Further research is very unlikely to change the authors' confidence in the estimated effect. Moderate quality: Further research is likely to have an important impact on confidence in the estimate of effect and may change the estimate.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The evidence from the buprenorphine meta-analysis (Mattick et al., 2014) derives from a search of databases from 2003 to 2013. Thirty-one randomized controlled trials of buprenorphine maintenance treatment versus placebo or methadone treatment for management of opioid use disorders were included.
Estimates of benefit and consistency across studies	Strong evidence indicated that buprenorphine is superior to placebo medication in retention of participants in treatment at all dosing levels considered in the analyses. Specifically, buprenorphine retained participants better than a placebo at low doses (2 to 6 mg), at medium doses (7 to 15 mg), and at high doses (≥ 16 mg). The authors based their conclusion on placebo-controlled trials, concluding that buprenorphine is an effective medication for retaining individuals with an OUD in treatment at any dose above 2 mg and for suppressing illicit opioid use (at doses 16 mg or greater).
What harms were identified?	No harms were identified.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See systematic reviews 3, 2, and 1, which support these recommendations.

Source of Systematic Review 5: • **Title** • **Author** • **Date**	**Source** • **Title:** Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence

• **Citation, including page number** • **URL**	• **Author:** Mattick RP, Breen C, Kimber J, and Davoli M. • **Date:** 2009 • **Citation:** Mattick RP, Breen C, Kimber J, and Davoli M. Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence. Cochrane Database of Systematic Reviews 2009, Issue 3. Art. No.: CD002209. • **URL:** http://onlinelibrary.wiley.com/doi/-10.1002/14651858.CD002209.pub2/epdf
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.	Another meta-analysis included 11 studies that met the criteria for inclusion in a Cochrane review of methadone treatment (R.P. Mattick et al., 2009). All the studies were randomized clinical trials, two were double-blind. Methadone appeared statistically significantly more effective than nonpharmacological approaches in retaining patients in treatment and in reducing heroin use as measured by self-report and urine/hair analysis. The authors concluded that methadone is an effective maintenance therapy intervention for the treatment of OUD as it retains patients in treatment and decreases heroin use better than treatments that do not use opioid replacement therapy. However, the authors did not show a statistically significant superior effect on criminal activity or mortality.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The Cochrane meta-analysis of methadone treatment (R.P. Mattick et al., 2009) also assigned high to moderate grades to the studies in the review.
Provide all other grades and definitions from the evidence grading system	High quality: Further research is very unlikely to change the authors' confidence in the estimated effect. Moderate quality: Further research is likely to have an important impact on confidence in the estimate of effect and may change the estimate.
Grade assigned to the **recommendation** with definition of the grade	The Cochrane meta-analysis of methadone treatment (R.P. Mattick et al., 2009) also assigned high to moderate grades to the studies in the review.
Provide all other grades and definitions from the recommendation grading system	High quality: Further research is very unlikely to change the authors' confidence in the estimated effect. Moderate quality: Further research is likely to have an important impact on confidence in the estimate of effect and may change the estimate.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The evidence from the methadone meta-analysis (Mattick et al., 2009) derives from a search of databases from 2001 to 2008. Eleven studies were included in the Cochrane Review (R.P. Mattick et al.,

	2009). The quality of the evidence varied from high to moderate. The results from the 11 randomized trials all showed statistically significant positive benefits from methadone treatment, despite small sample sizes. All the studies were randomized clinical trials, and two were double-blind.
Estimates of benefit and consistency across studies	Methadone treatment appeared to be statistically significantly more effective than nonpharmacological approaches in retaining patients in treatment and reducing opioid use as measured by self-report and urine/hair analysis (six RCTs, RR 0.66, 95 percent CI 0.56 to 0.78), but not statistically different in reducing criminal activity (three RCTs, RR 0.39, 95 percent CI 0.12 to 1.25) or mortality (four RCTs, RR 0.48, 95 percent CI 0.10 to 2.39).
What harms were identified?	Given that none of the negative effects described above were statistically significant differences, this study did not identify any harms.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See systematic reviews 4, 2, and 1, which support these recommendations.

Source of Systematic Review 6: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Source** • **Title:** Medication-Assisted Treatment for Opioid Addiction in Opioid Treatment Programs. Treatment Improvement Protocol (TIP) Series 43 • **Author:** Center for Substance Abuse Treatment • **Date:** 2005 • **Citation:** Center for Substance Abuse Treatment (2005). *Medication-Assisted Treatment for Opioid Addiction in Opioid Treatment Programs. Treatment Improvement Protocol (TIP) Series 43*. HHS Publication No. (SMA) 12-4214. Rockville, MD: Substance Abuse and Mental Health Services Administration. Page xvii. • **URL:** https://www.ncbi.nlm.nih.gov/books-/NBK64164/pdf/Bookshelf_NBK64164.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.	p. xvii of executive summary, “Research supports the perspective that opioid addiction is a medical disorder that can be treated effectively with medications when they are administered under conditions consistent with their pharmacological efficacy and when treatment includes necessary supportive services such as psychosocial counseling, treatment for co-occurring

	disorders, medical services, and vocational rehabilitation."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Not Applicable. The guidelines cited above do not provide grades (e.g., USPSTF grades A, B, etc.).
Provide all other grades and definitions from the evidence grading system	A consensus panel of experts developed the SAMHSA Treatment Improvement Protocols (CSAT TIP 43). A team of external field reviewers then reviewed and commented on the guideline recommendations.
Grade assigned to the **recommendation** with definition of the grade	Not Applicable. The guidelines cited above do not provide grades (e.g., USPSTF grades A, B, etc.).
Provide all other grades and definitions from the recommendation grading system	A consensus panel of experts developed the SAMHSA Treatment Improvement Protocols (CSAT TIP 43). A team of external field reviewers then reviewed and commented on the guideline recommendations.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The number of studies referenced in support of the guidelines was 608. The quality of the evidence selected indicates that included studies were considered to be of high quality by an expert panel.
Estimates of benefit and consistency across studies	Methadone maintenance is safe and effective, especially when used with psychosocial services (O'Connor & Fiellin, 2000). Maintenance treatment typically leads to reduction or cessation of illicit opioid use. A meta-analysis of 11 studies of the effectiveness of methadone (R.P. Mattick, Breen, Kimber, & Davoli, 2003) found that methadone was more effective than nonpharmacological treatment in retaining clients and reducing their opioid use. Clinical trials have demonstrated the primary efficacy of buprenorphine in patient retention as well as in the elimination of or reduction in opioid use (Fudala et al., 2003; Johnson, Strain, & Amass, 2003). Several studies evaluating the efficacy of buprenorphine for maintenance treatment lasting up to one year found that daily doses of 8 mg of sublingual solution or 8 to 16 mg of the buprenorphine tablet are safe and well tolerated. Most studies comparing buprenorphine and methadone have shown that 8 mg of sublingual buprenorphine or 16 mg of the tablet per day is equivalent to approximately 60 mg of oral methadone per day (Johnson et al., 2003). Naltrexone is effective in preventing relapse when used as directed; however, high rates of dropout have been reported. One study (Rothenberg et al., 2002) found especially poor retention among clients who

	had received methadone before naltrexone treatment. Naltrexone, under certain conditions, has resulted in better treatment compliance, e.g., when clients were supported with the opportunity to earn vouchers for treatment compliance (i.e., for each naltrexone dose ingested) (Preston et al., 1999). It should be noted, however, that CSAT TIP 43 predates approval of the long-acting naltrexone formulation for opioid use disorder and therefore does not address the effectiveness of the long-acting injectable preparation.
What harms were identified?	A small number of clients (10 percent) using naltrexone may experience gastrointestinal side effects that may necessitate their stopping the medication. Most clients, however, experience only mild, transient stomach upset. Some other side effects may include anxiety, nervousness, insomnia, headache, joint or muscle pain, and tiredness (Center for Substance Abuse Treatment, 2005).
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See below systematic reviews 5, 4, 3, 2, and 1, which support these recommendations.

Source of Systematic Review 7 • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Source** • **Title:** Clinical guidelines for the use of buprenorphine in the treatment of opioid addiction. Treatment Improvement Protocol (TIP) Series 40 • **Author:** Center for Substance Abuse Treatment • **Date:** 2004 • **Citation:** Center for Substance Abuse Treatment (2004). *Clinical guidelines for the use of buprenorphine in the treatment of opioid addiction. Treatment Improvement Protocol (TIP) Series 40*. DHHS Publication No. (SMA) 04-3939. Rockville, MD: Substance Abuse and Mental Health Services Administration. Page 50. **URL:** https://www.ncbi.nlm.nih.gov/books-/NBK64245/pdf/BookshelfNBK6445.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.	p. 50, “The consensus panel recommends that the buprenorphine/naloxone combination be used for induction treatment (and for stabilization and maintenance) for most patients [with an OUD].”

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Not Applicable. The guidelines cited above do not provide grades (e.g., USPSTF grades A, B, etc.).
Provide all other grades and definitions from the evidence grading system	A consensus panel of experts developed the SAMHSA Treatment Improvement Protocols (CSAT TIP 40). A team of external field reviewers then reviewed and commented on the guideline recommendations.
Grade assigned to the **recommendation** with definition of the grade	Not Applicable. The guidelines cited above do not provide grades (e.g., USPSTF grades A, B, etc.).
Provide all other grades and definitions from the recommendation grading system	A consensus panel of experts developed the SAMHSA Treatment Improvement Protocols (CSAT TIP 40). A team of external field reviewers then reviewed and commented on the guideline recommendations.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The number of studies referenced in support of the guidelines was 180. The quality of the evidence selected indicates that included studies were considered to be of high quality by an expert panel.
Estimates of benefit and consistency across studies	Clinical trials that compared buprenorphine to a placebo and to methadone have established the effectiveness of buprenorphine as a maintenance treatment of opioid addiction (Center for Substance Abuse Treatment, 2004). Buprenorphine treatment, compared to a placebo, is effective in reducing opioid use (Johnson et al., 1995). Evidence demonstrates that higher doses of buprenorphine and methadone are more effective in reducing opioid use (Ling et al., 1998; Petitjean et al., 2001; Schottenfeld, Pakes, Oliveto, Ziedonis, & Kosten, 1997). Another randomized trial found buprenorphine to be as effective as methadone, 60 mg/d, for retaining patients and reducing their opioid use. Both medications were superior to methadone at a lower level (20 mg/d) in reducing illicit opioid use and maintaining patients in treatment for 25 weeks (Johnson, Jaffe, & Fudala, 1992). A multisite office-based randomized study compared the effectiveness and safety of buprenorphine (16 mg) in combination with naloxone (4 mg), buprenorphine alone (16 mg), and a placebo (Fudala et al., 2003). The buprenorphine/naloxone in combination and buprenorphine alone demonstrated greater efficacy than the placebo. The proportion of urine samples that were negative for opiates was greater in the combined-treatment and buprenorphine-alone groups (17.8 and 20.7 percent, respectively) than in the placebo group (5.8 percent, $p < 0.001$ for both comparisons). Both buprenorphine treatment groups

	also reported significantly less opiate craving ($p <$ 0.001 for both comparisons with placebo).
What harms were identified?	Buprenorphine and combinations of buprenorphine and naloxone are generally well tolerated, although side effects reported with these medications include headache, anxiety, constipation, perspiration, fluid retention in lower extremities, urinary hesitancy, and sleep disturbance (Center for Substance Abuse Treatment, 2004).
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	See below systematic reviews 6, 4, 3, and 1, which support these recommendations.

_____

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.

References

American Society of Addiction Medicine. (2015). National practice guideline for the use of medications in the treatment of addiction involving opioid use. Retrieved from http://www.asam.org/quality-practice/guidelines-and-consensus-documents/npg/jam-article

Ayanga, D., Shorter, D., & Kosten, T. R. (2016). Update on pharmacotherapy for treatment of opioid use disorder. *Expert opinion on pharmacotherapy, 17*(17), 2307-2318.

Ball, J. C., & Ross, A. (1991). *The Effectiveness of Methadone Maintenance Treatment: Patients,Programs, Services, and Outcome*. New York: Springer Verlag.

Center for Substance Abuse Treatment. (2004). *Clinical guidelines for the use of buprenorphine in the treatment of opioid addiction. Treatment Improvement Protocol (TIP) Series 40. DHHS Publication No. (SMA) 04-3939.* . Retrieved from Rockville, MD: Substance Abuse and Mental Health Services Administration: https://www.ncbi.nlm.nih.gov/books/NBK64245/

Center for Substance Abuse Treatment. (2005). *Medication-Assisted Treatment for Opioid Addiction in Opioid Treatment Programs. Treatment Improvement Protocol (TIP) Series 43*. Retrieved from HHS Publication No. (SMA) 12-4214. Rockville, MD: Substance Abuse and Mental Health Services Administration:

Clark, R. E., Baxter, J. D., Aweh, G., O'Connell, E., Fisher, W. H., & Barton, B. A. (2015). Risk Factors for Relapse and Higher Costs Among Medicaid Members with Opioid Dependence or Abuse: Opioid Agonists, Comorbidities, and Treatment History. *J Subst Abuse Treat, 57*, 75-80. doi:10.1016/j.jsat.2015.05.001

Clark, R. E., Baxter, J. D., Barton, B. A., Aweh, G., O'Connell, E., & Fisher, W. H. (2014). The impact of prior authorization on buprenorphine dose, relapse rates, and cost for Massachusetts Medicaid

beneficiaries with opioid dependence. *Health Serv Res, 49*(6), 1964-1979. doi:10.1111/1475-6773.12201

Fudala, P. J., Bridge, T. P., Herbert, S., Williford, W. O., Chiang, C. N., Jones, K., . . . Buprenorphine/Naloxone Collaborative Study, G. (2003). Office-based treatment of opiate addiction with a sublingual-tablet formulation of buprenorphine and naloxone. *N Engl J Med, 349*(10), 949-958. doi:10.1056/NEJMoa022164

Fullerton, C. A., Kim, M., Thomas, C. P., Lyman, D. R., Montejano, L. B., Dougherty, R. H., . . . Delphin-Rittmon, M. E. (2014). Medication-assisted treatment with methadone: assessing the evidence. *Psychiatr Serv, 65*(2), 146-157. doi:10.1176/appi.ps.201300235

Johnson, R. E., Eissenberg, T., Stitzer, M. L., Strain, E. C., Liebson, I. A., & Bigelow, G. E. (1995). A placebo controlled clinical trial of buprenorphine as a treatment for opioid dependence. *Drug Alcohol Depend, 40*(1), 17-25.

Johnson, R. E., Jaffe, J. H., & Fudala, P. J. (1992). A controlled trial of buprenorphine treatment for opioid dependence. *JAMA, 267*(20), 2750-2755.

Johnson, R. E., Strain, E. C., & Amass, L. (2003). Buprenorphine: how to use it right. *Drug Alcohol Depend, 70*(2 Suppl), S59-77.

Ling, W., Charuvastra, C., Collins, J. F., Batki, S., Brown, L. S., Jr., Kintaudi, P., . . . Segal, D. (1998). Buprenorphine maintenance treatment of opiate dependence: a multicenter, randomized clinical trial. *Addiction, 93*(4), 475-486.

Mattick, R. P., Breen, C., Kimber, J., & Davoli, M. (2003). Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence (Cochrane Review). *Cochrane Database Systems Review 2003(2):CD00209.*

Mattick, R. P., Breen, C., Kimber, J., & Davoli, M. (2009). *Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence. Cochrane Database of Systematic Reviews 2009, Issue 3. Art. No.: CD002209. DOI: 10.1002/14651858.CD002209.pub2.* Retrieved from

Mattick, R. P., Breen, C., Kimber, J., & Davoli, M. (2014). *Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence*. Retrieved from

Mohlman, M. K., Tanzman, B., Finison, K., Pinette, M., & Jones, C. (2016). Impact of Medication-Assisted Treatment for Opioid Addiction on Medicaid Expenditures and Health Services Utilization Rates in Vermont. *J Subst Abuse Treat, 67*, 9-14. doi:10.1016/j.jsat.2016.05.002

O'Connor, P. G., & Fiellin, D. A. (2000). Pharmacologic treatment of heroin-dependent patients. *Ann Intern Med, 133*(1), 40-54.

Parran, T. V., Adelman, C. A., Merkin, B., Pagano, M. E., Defranco, R., Ionescu, R. A., & Mace, A. G. (2010). Long-term outcomes of office-based buprenorphine/naloxone maintenance therapy. *Drug Alcohol Depend, 106*(1), 56-60. doi:10.1016/j.drugalcdep.2009.07.013

Petitjean, S., Stohler, R., Deglon, J. J., Livoti, S., Waldvogel, D., Uehlinger, C., & Ladewig, D. (2001). Double-blind randomized trial of buprenorphine and methadone in opiate dependence. *Drug Alcohol Depend, 62*(1), 97-104.

Pierce, M., Bird, S. M., Hickman, M., Marsden, J., Dunn, G., Jones, A., & Millar, T. (2016). Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. *Addiction, 111*(2), 298-308. doi:10.1111/add.13193

Preston, K. L., Silverman, K., Umbricht, A., DeJesus, A., Montoya, I. D., & Schuster, C. R. (1999). Improvement in naltrexone treatment compliance with contingency management. *Drug Alcohol Depend, 54*(2), 127-135.

Rothenberg, J. L., Sullivan, M. A., Church, S. H., Seracini, A., Collins, E., Kleber, H. D., & Nunes, E. V. (2002). Behavioral naltrexone therapy: an integrated treatment for opiate dependence. *J Subst Abuse Treat, 23*(4), 351-360.

Schottenfeld, R. S., Pakes, J. R., Oliveto, A., Ziedonis, D., & Kosten, T. R. (1997). Buprenorphine vs methadone maintenance treatment for concurrent opioid dependence and cocaine abuse. *Arch Gen Psychiatry, 54*(8), 713-720.

Strain, E. C., Bigelow, G. E., Liebson, I. A., & Stitzer, M. L. (1999). Moderate- vs high-dose methadone in the treatment of opioid dependence: a randomized trial. *JAMA, 281*(11), 1000-1005.

Substance Abuse and Mental Health Services Administration. (2015). *Drug Addiction Treatment Act of 2000 (DATA 2000).* Retrieved from https://www.samhsa.gov/medication-assisted-treatment/legislation-regulations-guidelines#DATA-2000

Thomas, C. P., Fullerton, C. A., Kim, M., Montejano, L., Lyman, D. R., Dougherty, R. H., . . . Delphin-Rittmon, M. E. (2014). Medication-assisted treatment with buprenorphine: assessing the evidence. *Psychiatr Serv, 65*(2), 158-170. doi:10.1176/appi.ps.201300256

Measure Information

This document contains the information submitted by measure developers/stewards, but is organized according to NQF's measure evaluation criteria and process. The item numbers refer to those in the submission form but may be in a slightly different order here. In general, the item numbers also reference the related criteria (e.g., item 1b.1 relates to sub criterion 1b).

Brief Measure Information

NQF #: 3400
Corresponding Measures:
De.2. **Measure Title:** Use of pharmacotherapy for opioid use disorder (OUD)
Co.1.1. **Measure Steward:** Centers for Medicare & Medicaid Services, Centers for Medicaid & CHIP Services
De.3. **Brief Description of Measure:** The percentage of Medicaid beneficiaries ages 18 to 64 with an OUD who filled a prescription for or were administered or ordered an FDA-approved medication for the disorder during the measure year. The measure will report any medications used in medication-assisted treatment of opioid dependence and addiction and four separate rates representing the following types of FDA-approved drug products: buprenorphine; oral naltrexone; long-acting, injectable naltrexone; and methadone.
1b.1. **Developer Rationale:** Of the 52,404 drug overdose deaths in the United States in 2015, 33,091 (63.1 percent) were due to opioid use (Rudd, Seth, David, & Scholl, 2016) and an estimated 2.5 million individuals have an OUD for abuse or dependence with most not receiving treatment or not receiving the most effective care (Substance Abuse and Mental Health Services Administration, 2015). Among the outcomes that may be affected by OUD treatment are a reduction in drug use, medical problems, and criminal activity and improvements in vocational skills, employment, family relationships, and social activities (Center for Substance Abuse Treatment, 2005). Implementation of new treatment models to expand OUD treatment have been shown to be effective in increasing treatment capacity which is expected to influence patient outcomes (Brooklyn & Sigmon, 2017; Stoller, 2015). It is envisioned that the use of the measure, Use of Pharmacotherapy For Opioid Use Disorder, will improve quality of care by increasing the rate of pharmacotherapy among individuals with an OUD.

There is evidence that pharmacotherapy is related to improved outcomes, therefore, a quality measure to increase access to pharmacotherapy is expected to yield better care for beneficiaries with an OUD. Staying in methadone treatment has been associated with a reduced risk of death (Cousins et al., 2016). Several studies have shown that methadone is safe and effective, especially when higher doses (= 80mg/day) are provided (American Psychiatric Association, 2010). A meta-analysis that reviewed 11 studies on the effectiveness of methadone (Mattick, Breen, Kimber, & Davoli, 2003) found that methadone treatment was more effective than nonpharmacological treatment in retaining clients and reducing their opioid use. Another meta-analysis that reviewed 7 randomized controlled trials and 2 quasi-experimental studies of methadone maintenance found a high level of evidence that methadone treatment had a positive impact on retention in treatment and reduction in opioid use (Fullerton et al., 2014).

Sufficient evidence points to the safety and efficacy of buprenorphine for the treatment of OUD (Parran et al., 2010). The risk of fatal overdose on buprenorphine is substantially lower than that associated with the use of other opioid medications such as methadone because of the ceiling effects of buprenorphine across a wide range of doses (American Society of Addiction Medicine, 2015). One study found that buprenorphine at higher

doses (16 to 31mg) is as effective as methadone in reducing opioid use and improve treatment retention (Thomas et al., 2014).

A 2006 Cochrane review and 2009 update found oral naltrexone maintenance therapy alone or associated with psychosocial therapy to be more efficacious than placebo alone or associated with psychosocial therapy in limiting the use of heroin during the treatment, but not in improving retention, or preventing relapse (Minozzi et al., 2006). While oral naltrexone remains an FDA-approved medication for OUD, it has not been widely used due to concerns about adherence(ASAM Practice Guidelines, 2015) and need to maintain withdrawal prior to use (Center for Substance Abuse Treatment, 2005)

In a 6-month multisite double-blind, placebo-controlled RCT conducted in Russia, extended release naltrexone (XR-NTX) was found to be more efficacious than oral NTX with respect to treatment retention and reduction in use of illicit opioids (Krupitsky et al., 2012) and in a 1-year open-label extension of the original trial, about 51% of those who completed the extension were abstinent from opioids at all assessments during the 1-year open-label phase (Krupitsky et al., 2013). XR-NTX was also found to be effective in promoting abstinence across a range of demographic and baseline severity characteristics (Nunes et al., 2015).

S.4. **Numerator Statement:** Beneficiaries ages 18 to 64 with an OUD who filled a prescription for or were administered or ordered an FDA-approved medication for the disorder during the measure year.
S.6. **Denominator Statement:** Number of Medicaid beneficiaries with at least one encounter with a diagnosis of opioid abuse, dependence, or remission (primary or other) at any time during the measurement year.
S.8. **Denominator Exclusions:** None.

De.1. **Measure Type:** Process
S.17. **Data Source:** Claims
S.20. **Level of Analysis:** Population : Regional and State

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Not applicable; this measure is not a paired or grouped measure.

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

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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
NQF_3400__Evidence_Attachment.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.

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.

Of the 52,404 drug overdose deaths in the United States in 2015, 33,091 (63.1 percent) were due to opioid use (Rudd, Seth, David, & Scholl, 2016) and an estimated 2.5 million individuals have an OUD for abuse or dependence with most not receiving treatment or not receiving the most effective care (Substance Abuse and Mental Health Services Administration, 2015). Among the outcomes that may be affected by OUD treatment are a reduction in drug use, medical problems, and criminal activity and improvements in vocational skills, employment, family relationships, and social activities (Center for Substance Abuse Treatment, 2005). Implementation of new treatment models to expand OUD treatment have been shown to be effective in increasing treatment capacity which is expected to influence patient outcomes (Brooklyn & Sigmon, 2017; Stoller, 2015). It is envisioned that the use of the measure, Use of Pharmacotherapy For Opioid Use Disorder, will improve quality of care by increasing the rate of pharmacotherapy among individuals with an OUD.

There is evidence that pharmacotherapy is related to improved outcomes, therefore, a quality measure to increase access to pharmacotherapy is expected to yield better care for beneficiaries with an OUD. Staying in methadone treatment has been associated with a reduced risk of death (Cousins et al., 2016). Several studies have shown that methadone is safe and effective, especially when higher doses (= 80mg/day) are provided (American Psychiatric Association, 2010). A meta-analysis that reviewed 11 studies on the effectiveness of methadone (Mattick, Breen, Kimber, & Davoli, 2003) found that methadone treatment was more effective than nonpharmacological treatment in retaining clients and reducing their opioid use. Another meta-analysis that reviewed 7 randomized controlled trials and 2 quasi-experimental studies of methadone maintenance found a high level of evidence that methadone treatment had a positive impact on retention in treatment and reduction in opioid use (Fullerton et al., 2014).

Sufficient evidence points to the safety and efficacy of buprenorphine for the treatment of OUD (Parran et al., 2010). The risk of fatal overdose on buprenorphine is substantially lower than that associated with the use of other opioid medications such as methadone because of the ceiling effects of buprenorphine across a wide range of doses (American Society of Addiction Medicine, 2015). One study found that buprenorphine at higher doses (16 to 31mg) is as effective as methadone in reducing opioid use and improve treatment retention (Thomas et al., 2014).

A 2006 Cochrane review and 2009 update found oral naltrexone maintenance therapy alone or associated with psychosocial therapy to be more efficacious than placebo alone or associated with psychosocial therapy in limiting the use of heroin during the treatment, but not in improving retention, or preventing relapse (Minozzi et al., 2006). While oral naltrexone remains an FDA-approved medication for OUD, it has not been widely used due to concerns about adherence(ASAM Practice Guidelines, 2015) and need to maintain withdrawal prior to use (Center for Substance Abuse Treatment, 2005)

In a 6-month multisite double-blind, placebo-controlled RCT conducted in Russia, extended release naltrexone (XR-NTX) was found to be more efficacious than oral NTX with respect to treatment retention and reduction in use of illicit opioids (Krupitsky et al., 2012) and in a 1-year open-label extension of the original trial, about 51% of those who completed the extension were abstinent from opioids at all assessments during the 1-year open-

label phase (Krupitsky et al., 2013). XR-NTX was also found to be effective in promoting abstinence across a range of demographic and baseline severity characteristics (Nunes et al., 2015).

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.*

We tested the measure using 2014 Medicaid Analytic Extract data on 16 states. The number of beneficiaries with at least one opioid abuse, dependence, or in remission diagnosis varied across states (1,197 – 59,175). The overall performance rate for any pharmacotherapy use was 57.2%, and state-level scores ranged from 13.1% to 76.5%. Below we present state-specific performance rates. State-specific performance rates by four FDA-approved OUD medications are presented in the Testing Attachment (Table 6). State names are redacted

Use of Pharmacotherapy For Opioid Use Disorder, Rate Overall and By State
Overall rate across all states:
Numerator =116,593
Denominator = 203,816
rate =57.2

State A:
Numerator = 22,882
Denominator = 33,203
rate = 68.9
95% CI = 0.68,0.69

State B:
Numerator = 10,953
Denominator = 15,818
rate = 69.2
95% CI = 0.69,0.70

State C:
Numerator = 838
Denominator = 4,231
rate = 19.8
95% CI = 0.19,0.21

State D:
Numerator = 194
Denominator = 1,487
rate = 13.1
95% CI = 0.11,0.15

State E:
Numerator = 1,326
Denominator = 3,147
rate = 42.1
95% CI = 0.40,0.44

State F:
Numerator = 3,398

Denominator = 8,589
rate = 39.6
95% CI = 0.39,0.41

State G:
Numerator = 554
Denominator = 1,925
rate = 28.8
95% CI = 0.27,0.31

State H:
Numerator = 1,510
Denominator = 3,230
rate = 46.8
95% CI = 0.45,0.48

State I:
Numerator = 7,279
Denominator = 14,428
rate = 50.5
95% CI = 050,0.51

State J:
Numerator = 37,230
Denominator = 59,175
rate = 62.9
95% CI = 0.63,0.63

State K:
Numerator = 5,671
Denominator = 12,111
rate = 46.8
95% CI = 0.46,0.48

State L:
Numerator = 12,935
Denominator = 22,183
rate = 58.3
95% CI = 0.58,0.59

State M:
Numerator = 4,528
Denominator = 10,330
rate = 43.8
95% CI = 0.43,0.45

State N:
Numerator = 634
Denominator = 1,197
rate = 53.0
95% CI = 0.50,0.56

State O:
Numerator = 3,991
Denominator =5,217
rate = 76.5
95% CI = 0.75,0.78

State P:
Numerator = 2,670
Denominator = 7,545
rate = 35.4
95% CI = 0.34,0.36

Source: Based on analysis of 2014 MAX PS, IP, LT, OT, and RX files.
CI = confidence interval; NR=Not reported; result is based on a cell size of 10 or less.

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. We responded to 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.*
We used 2014 Medicaid Analytic Extract data on 16 states to test for disparities in performance rates by the following population groups: Medicaid beneficiary eligibility category, age, gender, race/ethnicity, and urban/rural. By conducting a Chi-squared test on each of these patient characteristics, we found significant variation in performance rates for each population group tested.

We found significant variation in the pharmacotherapy rate by beneficiary category. The adult beneficiary category had a 59.0 percent pharmacotherapy rate, the blind-disabled category had a 53.9 percent rate, the child category (which we assume includes the beneficiaries who are age 18) had a 34.7 percent rate, and the aged had a 74.3 percent rate.

Beneficiaries ages 45-64 had the highest overall pharmacotherapy rate (58.5 percent), followed closely by ages 25–44 (57.9 percent), whereas the younger age group 18–24 had the lowest rate (47.1 percent).

In the overall sample, females had a significantly lower pharmacotherapy rate than males (55.7 percent versus 58.9 percent); however, this overall trend is likely driven by similar relationship from several states with large number of OUD diagnosed beneficiaries.

In the overall sample, the beneficiaries of Hispanic/Latino (71.2 percent) or native Hawaiian/Pacific islander (73.8 percent) had the top pharmacotherapy rates, followed by white (57 percent), Asian (53.5 percent), and American Indian/Alaskan Native (52.8 percent). The black beneficiaries had the lowest rate (47.1 percent) among all major race groups.
Beneficiaries living in the urban area had a significantly higher pharmacotherapy rate (58.8 percent) than those who lived in the rural area (45.7 percent) overall.

In sum, testing results across all states showed higher pharmacotherapy rates for aged Medicaid beneficiaries, those who are older (age 25+), females, Latinos and Native Hawaiian /Pacific Islanders, and urban beneficiaries.

Below we show the number of beneficiaries with OUD diagnosis and the percent of beneficiaries with any OUD pharmacotherapy by beneficiary characteristics (rate).

Total
N=203,816
Rate=57.2

Medicaid beneficiary category
Aged
N=2,532
Rate = 74.3

Blind-disabled
N=71,170
Rate = 53.9

Adult
N=128,455
Rate= 59.0

Child
N=1,659
Rate = 34.7

Age
18–24
N=17,854
Rate = 47.1

25–44
N= 107,320
Rate = 57.9

45–64
N= 78,642
Rate = 58.5

Gender
Male
N= 97,668
Rate = 58.9

Female
N= 106,148
Rate = 55.7
Race.ethnicity
White
N= 125,416
Rate = 57.0

Black
N= 34,811
Rate = 47.1

American Indian/Alaskan Native
N= 1,306
Rate = 52.8

Asian
N= 1,155
Rate = 53.3

Hispanic/Latino
N= 32,141
Rate = 71.3

Native Hawaiian/Pacific Islander
N= 794
Rate = 73.8

Other race/ethnicity
N= 520
Rate = 25.6

Unknown race/ethnicity
N= 7,673
Rate = 49.3

Rural/urban
Rural
N= 25,234
Rate = 45.7

Urban
N= 178,340
Rate = 58.8

Unknown
N= 242
Rate = 47.5

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**
See 1b.4 above.

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:** NQF_Value_Sets_SUD-4_FINAL_SUD_team.01.24.18.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. This is a new 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).
Beneficiaries ages 18 to 64 with an OUD who filled a prescription for or were administered or ordered an FDA-approved medication for the disorder during the measure year.

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).
Beneficiaries identified as filling a prescription for or were administered or ordered an FDA-approved medication for OUD, during the 12-month measure year, through pharmacy claims (relevant NDC code) or through relevant HCPCS coding of medical service. Only formulations with an OUD indication (not pain management) are included in measure calculation.
The measure will be calculated both overall and stratified by four medications/mode of administration: buprenorphine; oral naltrexone; long-acting, injectable naltrexone; and methadone.
A list of value sets for the measure is attached in the Excel workbook provided for question S.2b. NDC codes listed are codes that were used in testing and are current as of June 2017.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Number of Medicaid beneficiaries with at least one encounter with a diagnosis of opioid abuse, dependence, or remission (primary or other) at any time during 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).
Medicaid beneficiaries age 18 through 64, enrolled for full 12 months of measurement year, and had at least one encounter with a diagnosis of opioid abuse, dependence, or remission (primary or other) at any time during the measurement year. ICD-9 and ICD-10 codes for OUD are provided in the attached Excel file in required format at S.2b.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
None.

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.)
The measure will be calculated both overall and stratified by four medications/mode of administration: buprenorphine; oral naltrexone; long-acting, injectable naltrexone; and methadone.
The NDC pharmacy codes used to identify the FDA-approved medications for OUD are listed in an Excel file attached in 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.)*
Step 1: Identify denominator
Identify Medicaid beneficiaries age 18 through 64 years with at least one encounter associated with a diagnosis of opioid abuse, dependence, or remission (primary or other diagnosis) during the measurement year and continuously enrolled during the measurement year. Age is calculated as of January 1 of the measurement year.

Step 2: Identify the numerator as beneficiaries with evidence of at least one prescription filled, or were administered or ordered an FDA-approved medication for the disorder during the measurement year.
The measure will report any medications used in MAT of opioid dependence and addiction and four separate rates representing the following types of FDA-approved drug products: buprenorphine; oral naltrexone; long-acting, injectable naltrexone; and methadone.

Step 2A: Identify beneficiaries with evidence of at least one prescription for buprenorphine at any point during the measurement year.

Step 2B: Identify beneficiaries with evidence of at least one prescription for oral naltrexone at any point during the measurement year.

Step 2C: Identify beneficiaries with evidence of at least one prescription for long-acting, injectable naltrexone at any point during the measurement year.

Step 2D: Identify beneficiaries with evidence of at least one prescription for methadone at any point during the measurement year.
Note: Pharmacotherapy for opioid abuse, dependence, or remission prescriptions and procedures, might occur in several files. Similarly, a diagnosis of opioid abuse, dependence, or remission might occur in several files. For example, one claims file may contain injectables while another claims file may contain oral medications. Consequently, pharmacotherapy and opioid abuse, dependence, or remission variables are created separately in each source and then merged by beneficiary ID.

Step 3: Calculate the overall rate by dividing the number of beneficiaries with evidence of at least one prescription (Step 2) by the number of beneficiaries with at least one encounter associated with a diagnosis of opioid abuse, dependence, or remission (Step 1). Then, calculate rates separately for each of the four medications.

Step 3A: Calculate the buprenorphine prescription rate by dividing the number of beneficiaries with evidence of at least one prescription for buprenorphine during the measurement year (Step 2A) by the number of beneficiaries with at least one encounter associated with a diagnosis of opioid abuse, dependence, or remission (Step 1).

Step 3B: Calculate the oral naltrexone prescription rate by dividing the number of beneficiaries with evidence of at least one prescription for oral naltrexone during the measurement year (Step 2B) by the number of beneficiaries with at least one encounter associated with a diagnosis of opioid abuse, dependence, or remission (Step 1).

Step 3C: Calculate the long-acting, injectable naltrexone prescription rate by dividing the number of beneficiaries with evidence of at least one prescription for injectable naltrexone during the measurement year (Step 2C) by the number of beneficiaries with at least one encounter associated with a diagnosis of opioid abuse, dependence, or remission (Step 1).

Step 3D: Calculate the methadone prescription rate by dividing the number of beneficiaries with evidence of at least one prescription for methadone during the measurement year (Step 2D) by the number of beneficiaries with at least one encounter associated with a diagnosis of opioid abuse, dependence, or remission (Step 1).

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 involve sampling.

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 or instrument.

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 Alpha-MAX 2014 data: 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.

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)*
Emergency Department and Services, 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
OUD__Pharmacotherapy_NQF_Testing_Attachment-636507481138452022.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.

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Use of pharmacotherapy for opioid use disorder (OUD)
Date of Submission: 1/5/2018
Type of Measure:

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

Instructions

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

Medicaid Analytic eXtract (MAX) 2014 eligible (EL), inpatient (IP), other services (OT), long-term care (LT) and drug (RX) files were used to conduct testing. 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..
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

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: State	☒ other: 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)

We included the following 16 states in measure testing. State names are redacted from this document.

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*)

Of the beneficiaries only eligible for Medicaid (and not both Medicaid and Medicare) and over 18 years in our sample states, 310,857 beneficiaries had at least one opioid abuse, dependence, or in remission diagnosis in calendar year 2014. Of these, 203,816 met the continuous enrollment requirement. Table 1 describes the beneficiaries included in the analytic sample. State names have been redacted.

Table 1. Analytic Sample Selection (1/1/2014 to 12/30/2014)

State	Total enrollment (age 18 and over) (N)	Total enrollment (age 18 and over) who are NOT dually eligible for Medicaid and Medicare or age ≥ 65 (N)	Medicaid beneficiaries with at least one opioid abuse, dependence, or in remission diagnosis during the year (N)	Medicaid beneficiaries with at least one opioid abuse, dependence, or in remission diagnosis during the year AND enrolled in Medicaid for 12 months (N)
Total	22,937,923	17,936,631	310,857	203,816
State A	10,000,779	8,400,335	46,208	33,203
State B	552,739	376,252	24,277	15,818
State C	808,995	480,540	6,003	4,231
State D	344,014	259,435	2,435	1,487
State E	610,160	393,786	4,565	3,147
State F	1,404,388	1,086,448	21,278	8,589
State G	358,314	190,243	2,433	1,925
State H	549,213	337,607	4,807	3,230
State I	985,551	747,206	23,275	14,428
State J	4,017,865	3,355,738	91,459	59,175
State K	717,397	593,201	14,491	12,111
State L	1,295,863	807,363	36,077	22,183
State M	650,183	424,925	12,203	10,330
State N	153,849	114,576	2,516	1,197
State O	120,563	88,534	6,615	5,217

State P	368,050	280,442	12,215	7,545

Two-thirds of the beneficiaries with an OUD diagnosis in the analytic sample were eligible for Medicaid under the "adult" eligibility category. Slightly more than half of the beneficiaries with an OUD diagnosis were ages 25 to 44 (Table 2), whereas 38.6 percent of beneficiaries with an OUD diagnosis were ages 45-64 years. Slightly more than half (52.1 percent) of beneficiaries with an OUD diagnosis were female. White beneficiaries accounted for almost two-thirds of beneficiaries with an OUD diagnosis (61.5 percent), followed by Black and Hispanic (17.1 and 15.8 percent, respectively). The testing and analyses included 203,816 beneficiaries with at least one OUD diagnosis during the year.

Table 2. Analytic Sample Demographic Information

Beneficiary characteristics	Number of beneficiaries with OUD diagnosis	Distribution of beneficiaries with OUD diagnosis (%)
TOTAL	**203,816**	**100.00**
Medicaid beneficiary category		
Aged	2,532	1.2
Blind-disabled	71,170	34.9
Adult	128,455	63.0
Child	1,659	0.8
Age		
18–24	17,854	8.8
25–44	107,320	52.7
45–64	78,642	38.6
Gender		
Male	97,668	47.9
Female	106,148	52.1
Race/ethnicity		
White	125,416	61.5
Black	34,811	17.1
American Indian/Alaskan Native	1,306	0.6
Asian	1,155	0.6
Hispanic/ Latino	32,141	15.8
Native Hawaiian/Pacific Islander	794	0.4
Other race/ethnicity	520	0.3
Unknown race/ethnicity	7,673	3.8

Beneficiary characteristics	Number of beneficiaries with OUD diagnosis	Distribution of beneficiaries with OUD diagnosis (%)
Rural/urban		
Rural	25,234	12.4
Urban	178,340	87.5
Unknown	242	0.1

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.

No difference in the data sample used for different aspects of testing.

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) 2014 files: Medicaid eligibility category, age, gender, and race/ethnicity, urban/rural, mental health status during the year, and SUD diagnosis during the year (other than OUD). This measure is based on a process that should be carried out for all beneficiaries, so no adjustment for patient mix is necessary.

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 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 for distinguishing 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-4 measure by first estimating the "noise" (within-plan variability), adjusted for the number of beneficiaries within that plan, and estimated the "signal" (between-plan variability). We computed the SNR statistic, R (Adams, 2009, 2014), as the ratio

of the signal variance (which is common across all entities) to the sum of the signal variance and the noise variance (which varies by entity):

$$R = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}$$

Temporal consistency. We assessed the temporal consistency (also referred to as temporal stability) of the SUD-4 measure by examining the strength of association between measure results in four quarters of the 2013 and 2014 measurement years. We then calculated Spearman's rank-order correlation coefficient (ranging from −1 to +1) between the measure results aggregated to the state level. 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. NQF typically considers an SNR statistic > 0.70 as acceptable for reliability (National Quality Forum, 2011). The SUD-4 was highly reliable in terms of ability to distinguish the measure's performance in different states, with an average reliability score of 0.998 across states and a range from 0.993 to 0.999 (Table 3). State names have been redacted.

Table 3. SUD-4 Measure rate and signal-to-noise reliability, by State

State	Number of eligible SUD-4 beneficiaries (denominator)	Number of beneficiaries receiving related pharmacotherapy for OUD (numerator)	SUD-4 measure rate	Signal-to-noise reliability
State A	33,203	22,882	0.689	0.999
State B	15,818	10,953	0.692	0.999
State C	4,231	838	0.198	0.999
State D	1,487	194	0.130	0.997
State E	3,147	1,326	0.421	0.997
State F	8,589	3,398	0.396	0.999
State G	1,925	554	0.288	0.996
State H	3,230	1,510	0.467	0.997
State I	14,428	7,279	0.505	0.999
State J	59,175	37,230	0.629	0.999
State K	12,111	5,671	0.468	0.999

State L	22,183	12,935	0.583	0.999
State M	10,330	4,528	0.438	0.999
State N	1,197	634	0.530	0.993
State O	5,217	3,991	0.765	0.999
State P	7,545	2,670	0.354	0.999

Notes: The signal-to-noise coefficients for State A, State B, State J and State L. Total were truncated to 0.999 rather than rounded to 1.000 to reflect the uncertainty in the estimates.

Note that high reliability is not indicative of high quality of health care, but rather indicates that the SUD-4 measure can be used to distinguish the measure's performance in different states. The high reliability for the measure at the state level is likely influenced by the adequate sample sizes, hence low "noise" variance.

Temporal consistency. As an indicator of reliability, the SUD-4 measure exhibits adequate stability over time. Specifically, the Spearman rank correlation of the state-level SUD-4 measure rate between CY 2013 and 2014 is 0.92 at the 95 percent confidence interval (0.77, 0.97). The high positive correlation indicates a strong tendency for the relative ranks of state-level measure rates to be stable over time.

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-4 is rated high for scientific acceptability, based on reliability testing results. Specifically, the high SNR indicated that the SUD-4 measure can discern the underlying performance between states within high precision. High temporal consistency showed that the performance of state-level SUD-4 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 <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **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)*

To investigate the convergent validity of the SUD-4 measure, we examined how state-level performance of SUD-4 compares to state-level performance on two Healthcare Effectiveness Data and Information Set (HEDIS) measures:

- **Initiation treatment (IT)**: Percentage of population between 18 and 64 years old with alcohol or drug (AOD) dependence who initiated treatment through an inpatient AOD admission, outpatient visit, intensive outpatient encounter, or partial hospitalization within 14 days of the diagnosis
- **Initiation and engagement treatment (IET)**: Percentage of population between 18 and 64 years old with Alcohol or Drug Dependence who initiated treatment and who had two or more additional services with a diagnosis of AOD within 30 days of the initiation visit

The data are from the Adult Health Care Quality Measures dataset, which includes performance rates on frequently reported health care quality measures in the CMS Medicaid/CHIP Adult Core Set of Behavioral Health measures (Centers for Medicare & Medicaid Services, 2017).

We assessed the convergent validity of the SUD-4 measure by calculating its Spearman rank correlation with the two HEDIS measures. The Spearman rank correlation ranges from −1 to 1, with positive value indicating a positive relation between the two measures and negative value showing an opposite direction of the two. Moreover, large magnitude (regardless of the sign) of the correlation value demonstrates a strong association between the two measures, whereas a correlation value close to zero implies a weak association.

To evaluate face validity, we surveyed a multi-stakeholder technical expert panel (TEP) that was convened to provide input and guidance on measure development activities under CMS contract HHSM-500-2013-13011I, Task Order # HHSM-500-T0004. The TEP includes 19 individuals representing consumers, state officials, health plans, provider organizations, researchers, and federal government agencies. We asked the TEP to rate their agreement that performance scores on the measure "Use of Pharmacotherapy for Opioid Use Disorder" can be used to distinguish good from poor quality of care. TEP members rated their agreement using a 4-point scale that ranged from strongly disagree to strongly agree.

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

Face validity. Nine out of 10 respondents agreed or strongly agreed that performance scores on the measure "Use of Pharmacotherapy for Opioid Use Disorder" can be used to distinguish good from poor quality of care.

Convergent validity. The state-level performances between SUD-4 and the IET measure appear to have a strong positive correlation. Specifically, we find states with high or low SUD-4 rates, respectively, in general tend to have high or low IET rates as well (Figure 1). The exception is State D, which has the lowest SUD-4 performance rate among all states, but has moderate IET performance. Moreover, the Spearman rank correlation between the SUD-4 measure and the IET

treatment initiation and engagement measure is 0.69, with the 95 percent confidence interval (0.20, 0.91). This indicates a strong correlation between the two measures.

Figure 1. Correlation between states' SUD-4 measure rates and IET measure rates

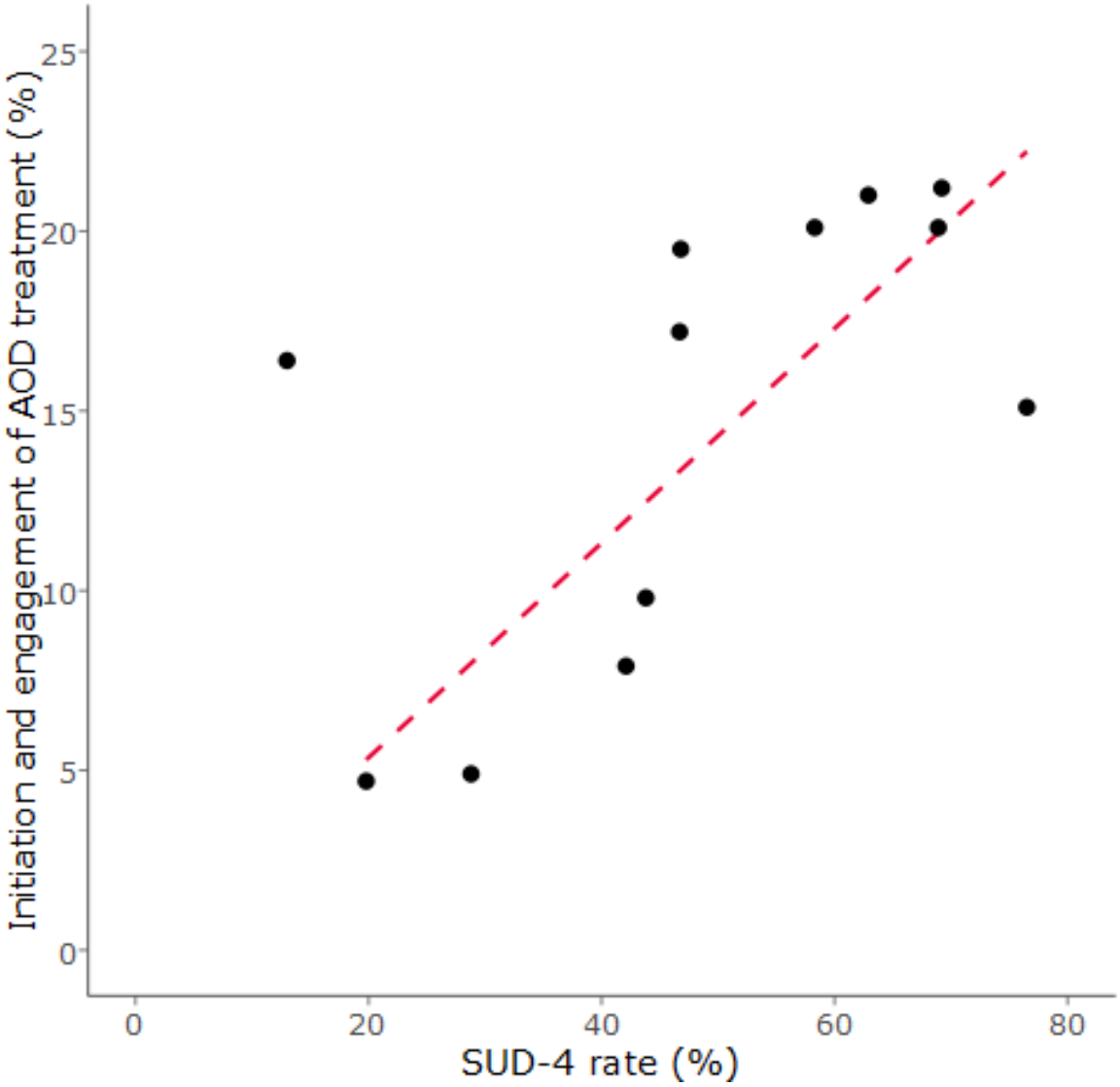


Table 4 shows the performance rates for three measures (SUD-4, initiation, and engagement) by state. Note that we have 16 states from our analytic data with SUD-4 measure rates, 12 of which also appear in the core set database for the IT and IET measures. Hence, our analysis below focuses on these 12 states. State names have been redacted.

Table 4. Performance rates for SUD-4, initiation, and initiation and engagement, by state

State	SUD-4	IT	IET
State A	68.9	36.2	20.1
State B	69.2	40.8	21.2
State C	19.8	35.1	4.7
State D	13.0	43.6	16.4
State E	42.1	36.9	7.9
State G	28.8	32.7	4.9
State H	46.7	44.3	17.2
State J	62.9	49.1	21
State K	46.8	39.1	19.5
State L	58.3	29.8	20.1
State M	43.8	37.2	9.8
State O	76.5	42.4	15.1

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 convergent validity of SUD-4 was excellent, showing states with high or low SUD-4 rates, respectively, in general tend to have high or low IET rates as well. Moreover, the Spearman rank correlation between the SUD-4 measure and the IET treatment initiation and engagement measure is 0.69, with the 95 percent confidence interval (0.20, 0.91). This indicates a strong correlation between the two measures.

2b2. EXCLUSIONS ANALYSIS

☒ **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.

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)*

We compared performance across state-level pharmacotherapy rates to understand any variation in performance. We calculated the 95% confidence interval of the pharmacotherapy rates 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. State measure rates that are significantly lower than the overall rate indicate an evidence of 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*)

Rates of pharmacotherapy varied by state. When considering beneficiaries with at least one medication of any type, State O, State B, and State A have the highest SUD-4 rates and State D, State C, and State G have the lowest (Table 5). As shown in Figure 2, all states are significantly different from the overall average. We found that the SUD-4 measure rates across the sixteen states cover a wide range with meaningful variation. Specifically, the measure rate ranges from 13.1 percent to 76.5 percent with mean of 57.2 percent and standard deviation of 13 percent. When looking at state-specific SUD-4 measure rates, 11 of the 16 states, or 68.8 percent, exhibit significantly lower measure rates than average, with their 95 percent CIs entirely below the overall performance rate. Five states show significantly higher measure rates than the average. Overall, the measure indicates both statistically significant and practically meaningful differences in performance across states.

Use of specific medications varies by state. Use of specific medications varied widely, from 35.9 percent for methadone and 21.8 percent for buprenorphine to the much lower frequency of 1.3 percent for oral naltrexone and 0.8 percent for injectable naltrexone (Table 6). We observe that five states had 0 beneficiaries with any methadone prescriptions (State D, State E, State G, State M, and State P), largely because methadone is not covered by Medicaid in these states. Most states have very few beneficiaries who received naltrexone (either oral or long acting

injectable), at less than 6 percent for all states combined, with most states' rates even lower. Since 2014, however, this distribution of medications may have changed, particularly with respect to greater use of injectable naltrexone.

Table 5. SUD-4 overall performance rate, by state

	Total beneficiaries with at least one opioid abuse, dependence, or in remission diagnosis	Percentage of beneficiaries with at least one opioid abuse, dependence, or in remission diagnosis with at least one medication (including methadone)		
		(N)	(%)	95% CI
Total	**203,816**	**116,593**	**57.21**	
State A	33,203	22,882	68.92	(0.68, 0.69)
State B	15,818	10,953	69.24	(0.69, 0.7)
State C	4,231	838	19.81	(0.19, 0.21)
State D	1,487	194	13.05	(0.11, 0.15)
State E	3,147	1,326	42.14	(0.4, 0.44)
State F	8,589	3,398	39.56	(0.39, 0.41)
State G	1,925	554	28.78	(0.27, 0.31)
State H	3,230	1,510	46.75	(0.45, 0.48)
State I	14,428	7,279	50.45	(0.5, 0.51)
State J	59,175	37,230	62.92	(0.63, 0.63)
State K	12,111	5,671	46.83	(0.46, 0.48)
State L	22,183	12,935	58.31	(0.58, 0.59)
State M	10,330	4,528	43.83	(0.43, 0.45)
State N	1,197	634	52.97	(0.5, 0.56)
State O	5,217	3,991	76.50	(0.75, 0.78)
State P	7,545	2,670	35.39	(0.34, 0.36)

Note: State names are redacted.

Source: Based on analysis of 2014 MAX PS, IP, LT, OT, and RX files.

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

Table 6. SUD-4 performance rate by medication and state

	Methadone[a]			Buprenorphine[a]			Naltrexone (oral)[a]			Naltrexone (injectable)[a]		
	(N)	(%)	95% CI	(N)	(%)	95% CI	(N)	(%)	95% CI	(N)	(%)	95% CI
Total	**73,144**	**35.89**		**44,426**	**21.80**		**2,686**	**1.32**		**1,603**	**0.79**	
State A	21,207	63.87	(0.63,0.64)	1,884	5.67	(0.05, 0.06)	53	0.16	(0,0)	11	0.03	(0,0)
State B	7,986	50.49	(0.5,0.51)	3,123	19.74	(0.19,0.2)	288	1.82	(0.02,0.02)	162	1.02	(0.01,0.01)
State C	157	3.71	(0.03,0.04)	671	15.86	(0.15,0.17)	20	0.47	(0,0.01)	0	0.00	(0,0)
State D	0	0.00	(0,0)	171	11.50	(0.1,0.13)	25	1.68	(0.01,0.02)	0	0.00	(0,0)
State E	0	0.00	(0,0)	1,284	40.80	(0.39,0.43)	52	1.65	(0.01,0.02)	21	0.67	(0,0.01)
State F	2,233	26.00	(0.25,0.27)	1,187	13.82	(0.13,0.15)	61	0.71	(0.01,0.01)	34	0.40	(0,0.01)
State G	0	0.00	(0,0)	554	28.78	(0.27,0.31)	NR	NR	NR	0	0.00	(0,0)
State H	511	15.82	(0.15,0.17)	913	28.27	(0.27,0.3)	143	4.43	(0.04,0.05)	96	2.97	(0.02,0.04)
State I	4,706	32.62	(0.32,0.33)	2,616	18.13	(0.18,0.19)	175	1.21	(0.01,0.01)	119	0.82	(0.01,0.01)
State J	23,399	39.54	(0.39,0.4)	14,636	24.73	(0.24,0.25)	948	1.60	(0.02,0.02)	235	0.40	(0,0)

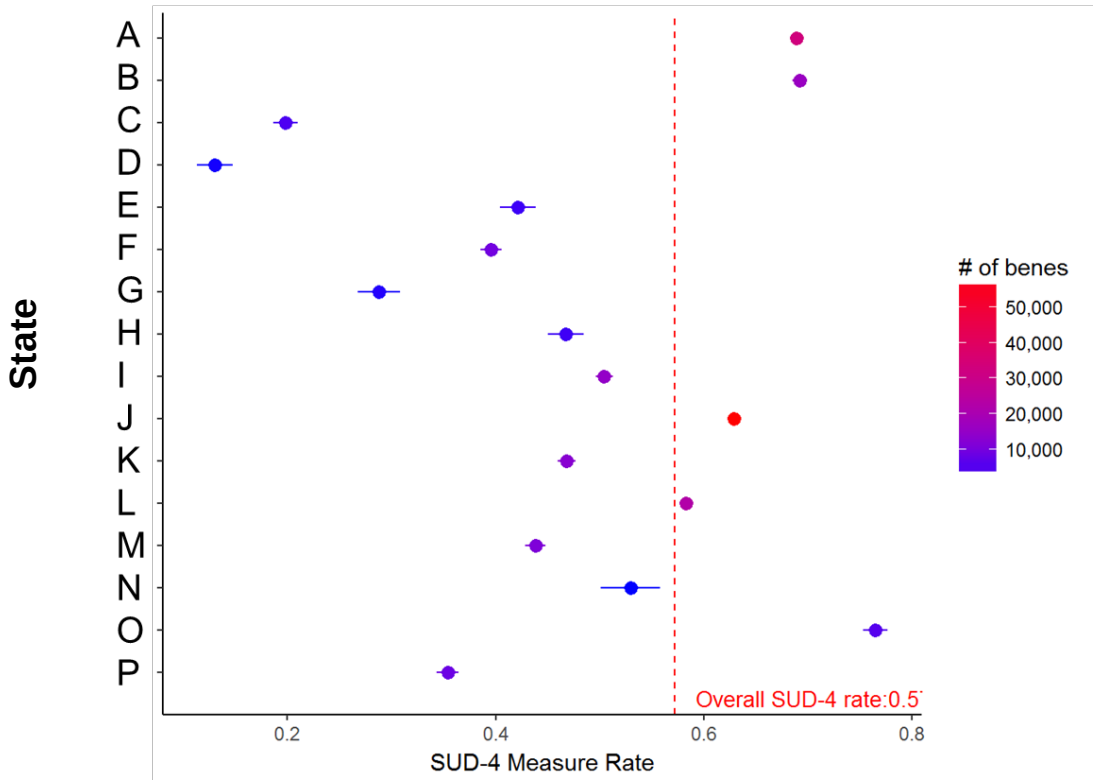
	Methadone[a]			Buprenorphine[a]			Naltrexone (oral)[a]			Naltrexone (injectable)[a]		
	(N)	(%)	95% CI	(N)	(%)	95% CI	(N)	(%)	95% CI	(N)	(%)	95% CI
State K	4,239	35.00	(0.34,0.36)	1,465	12.10	(0.12,0.13)	87	0.72	(0.01,0.01)	29	0.24	(0,0)
State L	6,396	28.83	(0.28,0.29)	6,860	30.92	(0.3,0.32)	411	1.85	(0.02,0.02)	218	0.98	(0.01,0.01)
State M	0	0.00	(0,0)	4,061	39.31	(0.38,0.4)	35	0.34	(0,0)	546	5.29	(0.05,0.06)
State N	396	33.08	(0.3,0.36)	240	20.05	(0.18,0.22)	18	1.50	(0.01,0.02)	16	1.34	(0.01,0.02)
State O	1,914	36.69	(0.35,0.38)	2,355	45.14	(0.44,0.46)	123	2.36	(0.02,0.03)	NR	NR	NR
State P	0	0.00	(0,0)	2,406	31.89	(0.31,0.33)	246	3.26	(0.03,0.04)	106	1.40	(0.01,0.02)

Source: Based on analysis of 2014 MAX PS, IP, LT, OT, and RX files.

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

Figure 2. SUD-4 Measure rate exhibits significant and clinically meaningful differences between states

A 95% confidence interval for the rate was calculated for each state, and compared to the overall rate, which is the mean state-level rate weighted by the number of beneficiaries in each state.



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 measure results suggest variation in performance and room for improvement in pharmacotherapy rates. When looking at state-specific SUD-4 measure rates, 11 of the 16 states, or 68.8 percent, exhibit significantly lower measure rates than average. Overall, the measure indicates both statistically significant and practically meaningful differences in performance across states.

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.***

Not applicable.

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*)

We assessed the extent of missing data 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-4 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-4 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-4 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 situation.
- Among the study states, the primary diagnosis code is mostly non-missing in the IP and LT files (Table 7). 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 58.9 percent in State M to 98.8 percent in State O (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 State J and State I, 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, State J 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 7. Percent of IP, LT, or OT file with primary diagnosis code or procedure code, and percent of RX file with days supply

	Percent with primary diagnosis code			Percent with procedure code	
State	**IP**	**LT**	**OT**	**IP**	**OT**
State A	100.0	100.0	74.3	64.9	86.2
State B	100.0	100.0	88.8	58.4	91.3
State C	100.0	100.0	95.7	60.5	96.3
State D	100.0	94.4	89.1	65.9	100.0
State E	100.0	89.1	88.0	62.8	99.3
State F	100.0	100.0	80.3	68.3	99.7
State G	100.0	100.0	83.9	31.8	99.1
State H	100.0	100.0	97.5	42.9	100.0
State I	100.0	100.0	97.4	69.2	96.7
State J	100.0	100.0	97.4	74.8	99.2

	Percent with primary diagnosis code			Percent with procedure code	
State	IP	LT	OT	IP	OT
State K	100.0	100.0	73.2	58.3	99.7
State L	100.0	100.0	97.3	67.4	100.0
State M	0.0	100.0	58.9	0.0	100.0
State N	100.0	100.0	74.5	51.5	99.8
State O	100.0	100.0	98.8	58.3	91.6
State P	100.0	100.0	90.7	59.7	98.9

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.

To calculate the SUD-4 measure generally and for specific subgroups, we also use data elements from the MAX PS file, including race, sex, age (calculated using date of birth), and eligibility information. Sex and date of birth are rarely missing (Table 8). Race, however, is missing for a substantial portion of enrollees in some states (for example, 43.8 percent of enrollees in State D), so examination of SUD-4 by race subgroup will exclude beneficiaries who are missing race data. Over 95 percent of MAX claims have corresponding Medicaid eligibility information (Table 9).

Table 8: Percent of Medicaid enrollees with missing date of birth, sex, or race

State	Percent of Enrollees Missing Date of Birth	Percent of Enrollees with Missing Sex	Percent of Enrollees with Missing Race
State A	0.0	0.0	64.5
State B	0.0	0.0	0.0
State C	0.0	0.0	11.1
State D	0.0	0.0	43.8
State E	0.0	0.0	8.9
State F	0.0	0.0	11.5
State G	0.0	0.0	6.1
State H	0.0	0.0	4.6
State I	0.0	0.0	28.3
State J	1.3	1.0	7.7
State K	0.0	0.0	19.7
State L	0.0	0.0	12.3
State M	0.0	0.0	10.9
State N	0.0	0.0	38.9

State O	0.0	0.0	26.2
State P	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.
Note: Numbers are from 2013 for all study states.

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)
State A	2013	2.62	0.46	0.05	0.80
	2014	1.70	0.22	0.02	0.64
State B	2013	0.27	0.22	0.07	0.18
State C	2013	0.96	0.12	0.02	0.17
	2014	0.85	0.07	0.01	0.16
State D	2013	0.17	0.14	0.04	0.01
	2014	0.08	0.04	0.02	0.00
State E	2013	4.65	1.48	0.27	3.18
State F	2013	4.08	1.59	0.41	0.38
	2014	1.50	0.94	0.46	0.10
State G	2013	2.03	0.14	0.01	0.97
	2014	0.35	0.21	0.02	0.07
State H	2013	0.11	0.54	0.02	0.04
	2014	0.23	0.28	0.02	0.10
State I	2013	0.55	0.20	0.42	0.21
	2014	0.55	0.24	0.33	0.21
State J	2013	0.07	0.23	0.21	0.00
State K	2013	0.35	0.02	0.01	0.04
State L	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
State M	2013	0.39	0.00	0.00	0.03
State N	2013	1.85	0.06	0.04	0.22
	2014	1.74	0.12	0.01	0.22
State O	2013	0.53	0.93	0.44	0.17
	2014	0.22	0.41	0.29	0.04
State P	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 of the study states in 2013. We have also provided 2014 information where available in the study states. State names are redacted.

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
- Centers for Medicare & Medicaid Services. (2017). 2017 Core Set of Behavioral Health Measures for Medicaid and CHIP (Behavioral Health Core Set). Retrieved from https://www.medicaid.gov/medicaid/quality-of-care/downloads/2017-bh-core-set.pdf
- National Quality Forum. (2011). Measure Testing Task Force Report. Retrieved from www.qualityforum.org/Publications/2011/01/Measure_Testing_Task_Force.aspx

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.

There are no 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)
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.*)
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 an OUD. Currently the overall rate of pharmacotherapy is 57.21% across the 16 states included in testing, and the range is 13.05% in State D to 76.59% in State O, indicating that there is an opportunity for improvement. The Use of Pharmacotherapy For Opioid Use Disorder measure may be useful for monitoring the rate of pharmacotherapy and encourage states to put interventions in place to increase the rates. This is important because pharmacotherapy has been shown to improve treatment retention and is related to better outcomes (Fullerton et al., 2014; Mattick et al., 2003; Minozzi et al., 2011; Thomas et al., 2014).

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 yet. There were no unexpected findings identified during testing of this measure.

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

This is a new measure that have not been implemented yet. No unexpected benefits were observed during testing.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

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

3175 : Continuity of Pharmacotherapy for Opioid Use Disorder

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

Evidence of medication-assisted treatment (MAT) among patients with opioid use disorder (OUD) or OD, Steward: OptumLabs

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.**

Not Applicable.

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 & 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:** Melissa, Azur, mazur@mathematica-mpr.com, 202-250-3518-

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's Technical Expert Panel provided input 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
- 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.

CPT(R) contained in the Measure specifications is copyright 2004-2016 American Medical Association.

ICD-10 copyright 2016 World Health Organization. All Rights Reserved.

The American Hospital Association holds a copyright to the National Uniform Billing Committee (NUBC) codes contained in the measure specifications. The NUBC codes in the specifications are included with the permission of the AHA. The NUBC codes contained in the specifications may be used by health plans and other health care delivery organizations for the purpose of calculating and reporting Measure results or using Measure results for their internal quality improvement purposes. All other uses of the NUBC codes require a license from the AHA. Anyone desiring to use the NUBC codes in a commercial product to generate measure results, or for any other commercial use, must obtain a commercial use license directly from the AHA. To inquire about licensing, contact ub04@healthforum.com.

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:** References

American Psychiatric Association. (2010). Practice Guidelines for the Treatment of Patients With Substance Use Disorders, second edition. Retrieved from https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/substanceuse.pdf

American Society of Addiction Medicine. (2015). National practice guideline for the use of medications in the treatment of addiction involving opioid use. Retrieved from http://www.asam.org/quality-practice/guidelines-and-consensus-documents/npg/jam-article

Brooklyn, J. R., & Sigmon, S. C. (2017). Vermont Hub-and-Spoke Model of Care For Opioid Use Disorder: Development, Implementation, and Impact. J Addict Med. doi:10.1097/ADM.0000000000000310

Center for Substance Abuse Treatment. (2005). Medication-Assisted Treatment for Opioid Addiction in Opioid Treatment Programs. Treatment Improvement Protocol (TIP) Series 43 HHS Publication No. (SMA) 12-4214. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2005, revised 2017.

Cousins, G., Boland, F., Courtney, B., Barry, J., Lyons, S., & Fahey, T. (2016). Risk of mortality on and off methadone substitution treatment in primary care: a national cohort study. Addiction, 111(1), 73-82. doi:10.1111/add.13087

Fullerton, C. A., Kim, M., Thomas, C. P., Lyman, D. R., Montejano, L. B., Dougherty, R. H., . . . Delphin-Rittmon, M. E. (2014). Medication-assisted treatment with methadone: assessing the evidence. Psychiatr Serv, 65(2), 146-157. doi:10.1176/appi.ps.201300235

Krupitsky, E., Nunes, E. V., Ling, W., Gastfriend, D. R., Memisoglu, A., & Silverman, B. L. (2013). Injectable extended-release naltrexone (XR-NTX) for opioid dependence: long-term safety and effectiveness. Addiction, 108(9), 1628-1637. doi:10.1111/add.12208

Krupitsky, E., Zvartau, E., Blokhina, E., Verbitskaya, E., Wahlgren, V., Tsoy-Podosenin, M., . . . Woody, G. E. (2012). Randomized trial of long-acting sustained-release naltrexone implant vs oral naltrexone or placebo for preventing relapse to opioid dependence. Arch Gen Psychiatry, 69(9), 973-981. doi:10.1001/archgenpsychiatry.2012.1a

Mattick, R. P., Breen, C., Kimber, J., & Davoli, M. (2003). Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence (Cochrane Review). Cochrane Database Systems Review 2003(2):CD00209.

Minozzi, S., Amato, L., Vecchi, S., Davoli, M., Kirchmayer, U., & Verster, A. (2006). Oral naltrexone maintenance treatment for opioid dependence. Cochrane Review 2006. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD001333.pub2/full

Minozzi, S., Amato, L., Vecchi, S., Davoli, M., Kirchmayer, U., & Verster, A. (2011). Oral naltrexone maintenance treatment for opioid dependence. Cochrane Database Syst Rev(4), CD001333. doi:10.1002/14651858.CD001333.pub4

Nunes, E. V., Krupitsky, E., Ling, W., Zummo, J., Memisoglu, A., Silverman, B. L., & Gastfriend, D. R. (2015). Treating Opioid Dependence With Injectable Extended-Release Naltrexone (XR-NTX): Who Will Respond? J Addict Med, 9(3), 238-243. doi:10.1097/ADM.0000000000000125

O´Connor, P. G., & Fiellin, D. A. (2000). Pharmacologic treatment of heroin-dependent patients. Ann Intern Med, 133(1), 40-54.

Parran, T. V., Adelman, C. A., Merkin, B., Pagano, M. E., Defranco, R., Ionescu, R. A., & Mace, A. G. (2010). Long-term outcomes of office-based buprenorphine/naloxone maintenance therapy. Drug Alcohol Depend, 106(1), 56-60. doi:10.1016/j.drugalcdep.2009.07.013

Rudd, R. A., Seth, P., David, F., & Scholl, L. (2016). Increases in Drug and Opioid-Involved Overdose Deaths — United States, 2010–2015. MMWR Morbity and Mortality Weekly Report, 65 1445–1452. DOI: http://dx.doi.org/1410.15585/mmwr.mm655051e655051.

Stoller, K. B. (2015). A collaborative opioid prescribing (CoOP) model linking opioid treatment programs with office-based burprenorphine providers. Addict Sci Clin Pract, 10 (Suppl 1), A63.

Substance Abuse and Mental Health Services Administration. (2015). Results from the 2014 National Survey on Drug Use and Health: Detailed Tables. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD. Table7.50A, accessed on February 23, 2017. Retrieved from https://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.pdf,

Thomas, C. P., Fullerton, C. A., Kim, M., Montejano, L., Lyman, D. R., Dougherty, R. H., . . . Delphin-Rittmon, M. E. (2014). Medication-assisted treatment with buprenorphine: assessing the evidence. Psychiatr Serv, 65(2), 158-170. doi:10.1176/appi.ps.201300256