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 #: 0008
Corresponding Measures:
Measure Title: Experience of Care and Health Outcomes (ECHO) Survey
Measure Steward: Agency for Healthcare Research and Quality
Brief Description of Measure: The ECHO is a survey that includes 5 multiple item measures and 12 single item measures:

Multiple Item Measures:

Getting treatment quickly
-Get treatment as soon as wanted when it was needed right away
-Get appointments as soon as wanted
-Get professional help by telephone
How well clinicians communicate
-Clinicians listen carefully
-Clinicians explain things in an understandable way
-Clinicians show respect
-Clinicians spend enough time
-Feel safe with clinicians
-Patient involved as much as wanted in treatment
Perceived improvement
-Compare ability to deal with daily problems to 1 year ago
-Compare ability to deal with social situations to 1 year ago
-Compare ability to accomplish things to 1 year ago
-Compare ability to deal with symptoms or problems to 1 year ago
Getting treatment and information from the plan
-Getting new clinician
-Delays in treatment while wait for plan approval
-Getting necessary treatment
-Understanding information about treatment in booklets or on the web
-Getting help when calling customer service
-Filling out paperwork
Informed about treatment options
-Told about self-help or consumer run programs
-Told about different treatments that are available for condition

Single Item Measures:

-Overall rating of counseling and treatment (MCO and MBHO)
-Overall rating of the health plan (MCO only)
-Wait more than 15 minutes past appointment time to see clinician
-Told about medication side effects
-Talk about including family & friends in treatment
-Given as much information as wanted about how to manage condition
-Given information about rights as a patient
-Patient feels that he or she could refuse a specific type of treatment
-Was information revealed that should have been kept private
-Cultural competence -Care responsive to language, race, religious, ethnic
-Amount helped by treatment
-Plan provides information about how to get treatment after benefits used up

The measures are based on reports of care experiences over the previous six months from adult (18 years of age or older) patients receiving behavioral health care (mental health and substance abuse treatment) and the organization that provides or manages their treatment and health outcomes.

Each measure score is the mean of the responses to the survey questions from patients receiving care at a particular health plan or managed behavioral health organization

More detail can be found at: http://www.ahrq.gov/cahps/surveys-guidance/echo/about/survey-measures.html
Developer Rationale: Donabedian and others have suggested that evaluating the process of care is one of the most direct ways to assess quality of care (Cleary, 2016). More recently, it has been suggested that one of eight ways that federal agencies can remove barriers to the delivery of effective behavioral care is measurement at the client, provider, organization, and population levels (Karakus, Ghose, et al.2016). The ECHO Survey assesses patient experiences with behavioral health services in such areas as getting treatment quickly, communication with clinicians, and information about treatment options. In moving away from global satisfaction questions, toward reports about specific well-defined aspects of care, the ECHO Survey more directly assesses quality of care than measures of "satisfaction".

The quality of behavioral health care is of significant concern because mental illnesses and alcohol and substance abuse impose substantial burdens on patients, employers, and the health care system. Payors and regulators are interested in assessing quality by comparing the experiences of patients receiving care from different health care organizations. A fundamental goal of CAHPS instruments is to provide evidence that can be used to improve health care and this was a major reason for developing the ECHO measures:

Shaul and colleagues noted that surveys such as the ECHO can identify aspects of the plan and treatment that are improvement priorities. Use of these data is likely to extend beyond the behavioral health plan to consumers, purchasers, regulators and policymakers, particularly since NCQA is encouraging behavioral health plans to use a similar survey for accreditation purposes.

To facilitate the use of such measures to stimulate quality improvement efforts, the CAHPS team reviewed the literature on strategies for improving aspects of care assessed in CAHPS surveys and developed: "The CAHPS Ambulatory Care Improvement Guide: Practical Strategies for Improving Patient Experience", which can be seen at:

http://www.ahrq.gov/cahps/quality-improvement/improvement-guide/improvement-guide.html:

Cleary, P. D. (2016). Evolving Concepts of Patient-Centered Care and the Assessment of Patient Care Experiences: Optimism and Opposition. Journal of health politics, policy and law, 3620881.

Karakus, M., Ghose, S. S., Goldman, H. H., Moran, G., & Hogan, M. F. (2016). "Big Eight" Recommendations for Improving the Effectiveness of the US Behavioral Health Care System. Psychiatric Services, appi-ps.

Shaul JA, Eisen SV, Stringfellow VL, Clarridge BR, Hermann RC, Nelson D, Anderson E, Kubrin AI, Leff HS, Cleary PD. Use of consumer ratings for quality improvement in behavioral health insurance plans. Jt Comm J of Qual Imp; 2001; 27: 216-229.

Numerator Statement: No changes form original specification: The ECHO survey measures patient-centered care by asking about patient experiences with behavioral health care (mental health and substance abuse treatment) and the organizations that provide or manage the person´s treatment and health outcomes.

The survey and instructions are available at:
www.qualityforum.org/pdf/ambulatory/txECHOALL(onepager&specs&survey)03-23-07.pdf
Measure developer/instrument web site:
www.cahps.ahrq.gov/content/products/ECHO/PROD_ECHO_MBHO.asp?p=1021&s=214
The composite measures' component items can be found on the document titled "Reporting Measures for the ECHO Survey 3.0" (Document No. 209 – 8/31/06) available for download at http://www.ahrq.gov/cahps/surveys-guidance/echo/instructions/index.htmlNo changes form original specification: The ECHO survey measures patient-centered care by asking about patient experiences with behavioral health care (mental health and substance abuse treatment) and the organizations that provide or manage the person´s treatment and health outcomes.

The survey and instructions are available at:
www.qualityforum.org/pdf/ambulatory/txECHOALL(onepager&specs&survey)03-23-07.pdf
Measure developer/instrument web site:
www.cahps.ahrq.gov/content/products/ECHO/PROD_ECHO_MBHO.asp?p=1021&s=214
The composite measures' component items can be found on the document titled "Reporting Measures for the ECHO Survey 3.0" (Document No. 209 – 8/31/06) available for download at http://www.ahrq.gov/cahps/surveys-guidance/echo/instructions/index.html

Denominator Statement: All survey respondents, or for selected items, all respondents who respond appropriately to screening questions.
Denominator Exclusions: No changes: Patients who received behavioral health services only in primary care settings (e.g. psychotropic medications from their primary care physician) are not included.

Measure Type: Outcome: PRO
Data Source: Patient Reported Data
Level of Analysis: Health Plan

Original Endorsement Date: Jul 01, 2007 **Most Recent Endorsement Date:** Jul 01, 2007

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 health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale. The evidence for a Patient-Reported Outcome-Based Performance Measures (PRO-PM) also should demonstrate that the target population values the measured PRO and finds it meaningful.

This submission contains information for 17 Patient-Reported Outcome based Performance Measures (PRO-PMs) that are calculated from data aggregated from responses to the Experience of Care and Health Outcomes (ECHO) survey. These 17 PRO-PMs include:

1) Getting treatment quickly
2) How well clinicians communicate
3) Perceived improvement
4) Getting treatment and information from the plan
5) Informed about treatment options
6) Overall rating of counseling and treatment (MCO and MBHO)
7) Overall rating of the health plan (MCO only)
8) Wait more than 15 minutes past appointment time to see clinician
9) Told about medication side effects
10) Talk about including family & friends in treatment
11) Given as much information as wanted about how to manage condition
12) Given information about rights as a patient
13) Patient feels that he or she could refuse a specific type of treatment
14) Was information revealed that should have been kept private
15) Cultural competence -Care responsive to language, race, religious, ethnic
16) Amount helped by treatment
17) Plan provides information about how to get treatment after benefits used up

NOTE that NQF's evaluation process for patient-reported outcome-based performance measures (PRO-PMs) has changed substantially since this measure was last endorsed (in 2007). The process now includes individual consideration of the PRO-PMs associated with a particular instrument and requires evidence, gap, and score-level reliability and validity testing for each included PRO-PM. When last endorsed, the Panel that evaluated the measure acknowledged several weaknesses of the ECHO survey itself, but agreed on the need for measures of patient experience for behavioral health.

Summary of prior review in 2007: N/A

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 a very basic logic model linking structures and process of care, generally, to patient experiences.

- The developer cites recent literature supporting a link between patient-centered care and other measures of healthcare quality. However, this linkage does not adequately address NQF's requirement to provide a rationale to support the relationship of the outcome (in this case, each of the 17 PRO-PMs under evaluation as part of this submission) and at least one healthcare structure, process, intervention, or service [*for example, behavioral health care providers might increase staff ratios or offer extended hours in order to improve patient perceptions regarding receipt of timely treatment*]. At least a couple of the cited articles (e.g., Anhang Price, Elliott, Cleary, et al. 2014; Wilson, et al., 2007) may address some of the outcomes of interest, but additional summarization is needed. Additional references cited by the developers also may include the necessary rationales; however, these must be summarized in the submission materials (citations of external materials are not sufficient).
- The developer does not provide evidence that the target population (i.e., those receiving behavioral health services) values the measured PROs and finds them meaningful.

Question for the Committee:

- *For each of the 17 PRO-PMs included in this submission, is there at least one thing that providers can do to achieve a change in the measure results?*
- *Are you aware of research that the outcomes assessed in this submission (quick treatment, good clinician communication, etc.) is valued by those receiving behavioral health services?*

Guidance from the Evidence Algorithm

Patient-reported outcome-based performance measures (Box 1) → no healthcare actions identified that can affect the PRO-PMs included in the measure (Box 2) → No Pass

The highest possible rating is PASS.

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

RATIONALE: Developer should identify at least one healthcare structure, process, intervention, or service that can affect patient experiences of behavioral healthcare for each of the 17 PRO-PMs included in the measure. Also, developer should demonstrate that those receiving behavioral health services find the various outcomes meaningful. NQF staff assume that these are available and would just need to be included explicitly in the submission materials.

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.

- No recent data on performance results were provided for the 17 PRO-PMs included in this submission.
- Some information on performance was provided for a few of the outcomes, but this appears to be limited to field testing conducted during development of the survey in the late 1990's.
- The developers also provide data from other patient experience PRO-PMs (i.e., Hospital CAHPS measures) and cite several articles to demonstrate disparities in access to services and treatment.

Disparities

- The developer summarizes two recent studies that used data derived from the ECHO survey.
  - Leff and colleagues (2016) found that four ethnically and racially diverse (ERD) groups differed in terms of preferred cultural elements but not in overall perceived quality of care. However, the developer does not describe which, if any, of the 17 experience-of-care PRO-PMs included in this submission were associated with race/ethnicity.
  - Martino and colleagues (2016) found that commercially-insured beneficiaries were more likely to have "*better general and mental health*" than those insured by Medicaid. However, It is not clear from the

summary provided which, if any, of the 17 experience-of-care PRO-PMs included in this submission were associated with insurance type.

Questions for the Committee:

- *Are you aware of any data that would demonstrate a gap in care or substantial variation in care among health plans for any of the 17 PRO-PMs included in this measure? If so, do those gaps warrant endorsement of a national performance measure?*
- *Are you aware of evidence that disparities exist for any of the 17 PRO-PMs included in this measure?*

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

RATIONALE: For measures undergoing maintenance evaluation, data on measure performance is required (data should be relatively recent).

Committee pre-evaluation comments

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

1.a. Evidence to Support Measure Focus

Comments:

**The developer does not provide sufficient information on how information gained from the ECHO survey can be used and has been used to improve services to consumers of behavioral health services. There is research and evaluations that suggest that involvement in treatment planning positively impacts perceptions of behavioral health care. The information provided by the surveys could be used to improve the treatment planning process, and the tracking of outcomes throughout the consumers involvement with the clinician.

**Specific evidence for each survey item (PRO) is not provided. Literature cited does support idea patient experience is linked to quality of care. No specifics regarding how the survey results could improve care are provided. The links to the survey cited in the documentation lead to non-existent web page.

**Some of these measures are in the control of or most related to actions of the healthcare provider and others more in control of or related to actions of the MCO/MBHO. I see information that could be garnered from these survey items that could possibly lead to corrective actions but, as stated in the preliminary analysis, the submission is not structured to meet the PRO-PM criteria that exists today and offers insufficient specific evidence.

1.b. Performance Gap

Comments:

** There was insufficient information provided by the developers regarding performance gaps related to the 17 Pro-PM measures. The relationship between the CABHS and ECHO surveys was only briefly touched upon in the application, and yet much of the information provided by the developers was related to CABHS. The developers indicated there was a lack of performance data on the ECHO survey. The information on the gaps was more general in nature, and not necessarily related to the 17 items.

**No recent data on performance results were provided for the 17 PRO-PMs included in this submission. Some information on performance was provided for a few of the outcomes, but this appears to be limited to field testing conducted during development of the survey in the late 1990's. The developers also provide data from other patient experience PRO-PMs (i.e., Hospital CAHPS measures) and cite several articles to demonstrate disparities in access to services and treatment.

Leff and colleagues (2016) found that four ethnically and racially diverse (ERD) groups differed in terms of preferred cultural elements but not in overall perceived quality of care. However, the developer does not describe which, if any, of the 17 experience-of-care PRO-PMs included in this submission were associated with race/ethnicity.

**I don't know of research specific to variation in care among plans or evidence of disparities tied to the specific PRO-PMs.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Self-reports of health plan enrollees who have received behavioral health care (mental health or substance abuse treatment)

Specifications:

- Each of the 17 PRO-PMs is specified at the health plan level of analysis, for use in the behavioral health outpatient setting. A higher score indicates better quality.
- Five of the PRO-PMs are based on multiple questions (or items) in the survey; 12 of the PRO-PMs are based on one question (or item) in the survey.
- Measure results are the case-mix-adjusted averages of the response(s) (across multiple items for the multi-item PRO-PMs and for the one item for the single-item PRO-PMs), aggregated across patients in a particular health plan.
- The target population for the ECHO survey is health plan members ages 18 years or older who have been continuously enrolled for 12 months and who have received outpatient behavioral health services. Diagnosis and procedural codes identifying "qualifying" behavioral health services have not been provided.
- Exclusions to the denominator include patients who received behavioral health services only in primary care settings.
- A brief calculation algorithm is provided, but it is not sufficient to describe how the measure results are calculated. It does, however, allude to a SAS "CAHPS Macro" that can be used to calculate measure results.
- The developer reports that the measure is case-mix (risk) adjusted for the following factors: self-reported mental health status, self-reported general health status, alcohol/drug use, age, education, and race/ethnicity. Presumably all 6 of these risk factors are included in the risk adjustment approach for all 17 of the PRO-PMs (although the coefficients for the risk factors would differ).
- No mention is made regarding mode of survey administration.
- Random sampling of eligible patients is allowed. Developers suggest fielding a large enough sample to ensure 411 completed surveys per health plan.
- Proxy responses are not allowed.

Questions for the Committee::

- *Is it clear which health plan beneficiaries would be eligible for the measure? Can these be consistently identified across health plans?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment

Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Internal consistency of the data elements was assessed, as were item-total correlations.

Describe any updates to testing: None reported.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☒ **No**

Method(s) of reliability testing

- Data used in testing included surveys from approximately 16,000 plan members from three states (New Jersey, Minnesota, and an unknown MBHO). These data represented responses from beneficiaries from fee-for-service, managed care, commercial health maintenance organizations, and public assistance health programs.
- Data element reliability for the 5 multi-item PRO-PMs was calculated by examining 1) the internal consistency using Cronbach's alpha (a measure of how well items "hang together" to measure the same underlying construct) and 2) the item-total correlation (the correlation between the item score and the multi-item score if that item is removed). These are appropriate methods for assessing data element reliability. ***No testing information for the single-item PRO-PMs was provided.***
- Using data from New Jersey, the developer reports testing for measure score reliability using one-way analysis of variance. In general, this is an appropriate method for assessing score-level reliability. However, the results provided indicate that this analysis was done for the items in the survey, rather than for the PRO-PM measure results. Therefore, this testing does not meet NQF's requirements for score-level testing of the PRO-PMs, which requires analysis to quantify the ability to detect differences in PRO-PM results between health plans.

Results of reliability testing

- Data element testing

Measure	Minnesota		MBHO		New Jersey	
	Coefficient alpha	Item-to-total correlation (range)	Coefficient alpha	Item-to-total correlation (range)	Coefficient alpha	Item-to-total correlation (range)
Getting treatment quickly	0.80 (n=600)	0.55 - 0.71	0.72 (n=162)	0.48 - 0.59	0.84 (n=458)	0.58 - 0.78
How well clinicians communicate	0.86 (n=1766)	0.56 - 0.71	0.85 (n=829)	0.54 - 0.70	0.89 (n=2154)	0.67 - 0.78
Perceived Improvement	0.83 (n=1769)	0.68 – 0.70	0.89 (n=837)	0.75 - 0.78	0.89 (n=2102)	0.73 - 0.77
Getting treatment and information from the MBHO or Plan	0.80 (n=68)	0.41 – 0.75	0.75 (n=294)	0.60	0.84 (n=136)	0.55 - 0.66
Informed about treatment options	0.66 (n=1766)	0.49	0.65 (n=818)	0.48	NA	NA

- For Cronbach's alpha, 0.70 or higher is a widely-accepted rule of thumb for a set of items to be considered a scale.
  - For 4 of the 5 multi-item scales, Chronbach's Alpha values ranged from 0.72 to 0.89. Reliability was lower for the *Informed about Treatment Options* scale. Item-to-total correlation values were quite high for all 5 scales, suggesting utility of the individual items in the scales.

Questions for the Committee:

- *Are the test samples used for the data element testing adequate to ensure reliability of the multi-item scales in the instrument?*

Guidance from the Reliability Algorithm

Specifications are mostly precise/complete, although detail on how to identify patients in the denominator is not provided (Box 1) → Score-level testing at the plan level for each PRO-PM was not provided, although this is a requirement for PRO-PMs (Box 3) → Insufficient

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

RATIONALE: Analysis to demonstrate ability to differentiate between measured entities is required for all PRO-PMs. Even though some plan-level testing was conducted, it was for items in the survey, not for the computed measure score(s). Data element level testing information (i.e., to demonstrate reliability of the survey items) also is required, but was provided only for the multi-item scales.

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☐ **Somewhat** ☒ **No**

Specification not completely consistent with evidence: As noted earlier, the submission materials are insufficient for the evidence subcriterion.

2b2. Validity testing

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

For maintenance measures, summarize the validity testing from the prior review: For the ECHO survey: content validity, cognitive testing to validate the questions used, and construct validity

Describe any updates to validity testing:

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing ~~against a gold standard~~** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- Data element testing
  - Three focus groups were conducted to obtain information regarding the content of the ECHO survey. Participants included those with experience in providing a variety of mental health services.
  - Cognitive testing to ensure understandability of the questions was conducted, although it is not clear how this was done or who was asked to participate.
- Construct validation was conducted by correlating responses from the multi-item scales to those of the treatment rating and plan rating items (which were presumably used as criterion variables). Results from the remaining 10 single items also were correlated with the treatment rating and plan rating items. The data used in testing included surveys from approximately 16,000 plan members from three states (New Jersey, Minnesota, and an unknown MBHO).
  - No explanation of this testing was provided, so it is not clear whether this represents validation of the survey items (i.e., data element validation) **or** validation of the plan-level results, where the results are

calculated as specified in this submission. NOTE that the latter is required for all PRO-PMs, and therefore, further clarification from the developers will be needed.

Validity testing results:

- Data element testing
  - Content validity: The developer noted that topic areas not initially assessed (e.g., adequacy of insurance coverage) was identified by the focus groups.
  - Cognitive testing: The developer states that testing ensured "*that questions are consistently understood and that questions are understood in the manner intended by questionnaire developers*".
- Construct validation: In general, for the 3 "state" samples, all 5 of the multi-item scales (or PRO-PMs) and most of the single items (or PRO-PMs) were positively and statistically significantly correlated with the two global items (or PRO-PMs). The items (or PRO-PMs) on including family and friends and refusing treatment were not significantly correlated with the plan rating item (or PRO-PM).

Questions for the Committee:

- *Based on the survey validation results that were presented, do you agree that the items included in the ECHO survey are valid?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients who received behavioral health services only in primary care settings are excluded from the measure.
- The developer does not provide any information about how many patients are actually excluded from the measure.

Questions for the Committee:

- *Is the validity of the measure threatened by excluding patients who receive behavioral health services in primary care settings only?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

- Note that the developers checked the box for risk-adjustment via stratification, but the description suggests a statistical modeling approach. The developers also indicated use of separate risk model for Commercial and Medicaid plans.

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**

SDS factors included in risk model? ☒ **Yes** ☐ **No**

Risk adjustment summary

- Presumably, each of the 17 PRO-PM measures are case-mix adjusted using a linear regression model [***this must be verified by the developer, especially given the submission materials explicitly mention only 4 of the 5 multi-item PRO-PMs and 2 of the single-item PRO-PMs***]
- Data used to develop the risk-adjustment approach included information from the ECHO field testing and survey data submitted to the National CAHPS Benchmarking Database. These data included responses from 4,068 health plan enrollees from Minnesota, New Jersey, Colorado, Florida, New York, and Ohio.
- Presumably, the same 6 factors are included in the case-mix adjustment approach for each of the 17 PRO-PMs. These include:
  - self-reported mental health status

- self-reported general health status
- alcohol/drug use
- age
- education
- race/ethnicity

- The final case-mix factor parameters for the 17 PRO-PMs have not been provided.
- It is not clear whether the case-mix adjustment models have or will be re-calibrated on a routine basis.
- No mention is made regarding adjustment (if any) based on mode of administration.

Conceptual analysis of the need for SDS adjustment:

- The developer noted that potential risk factors were those associated with consumer rating of behavioral health plans. However, they did not amplify on how SDS factors are related to consumer ratings or why these relationships might exist (e.g., why/how different racial/ethnic groups rate behavioral health plans differently).

Empirical analysis of SDS factors:

- The developers identify several analyses they conducted for both SDS and non-SDS factors in their case-mix model-building approach.
- It does not appear that a comparison was made to determine effects of risk-adjustment with inclusion of SDS factors versus without inclusion of SDS factors (although this is a requirement of NQF's SDS Trial).

Risk Model Diagnostics:

- The developers do not provide typical model discrimination or calibration statistics to describe the "goodness of fit" of the risk-adjustment approach (e.g., R^2 values, risk-decile plots)
- Developers noted that the effects of risk-adjustment "*generally appear to be modest*".

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included for each of the 17 PRO-PMs?*
- *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, describe the rationale provided.*

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

- Although the developer notes that "*most of the ECHO Survey items are able to discriminate between behavioral health plans*", this does not necessarily speak to whether there are differences between health plans based on the results of the 17 PRO-PMs, when calculated as specified in this submission. NOTE that if the results alluded to do reflect plan-level results, additional information will be needed to understand which of the 8 were unable to differentiate between plans.

2b6. Comparability of data sources/methods:

- Not applicable.

2b7. Missing Data

- The developer indicates missing data due to non-response is addressed via guidance on how to increase the number of returned surveys. They also report response rates of ranging from 36% to 65%, depending on type of plan (commercial vs. Medicaid) and depending on mode of administration. NOTE, however, that these response data are based on fielding of the survey in the late 1990's and may not reflect current experience.
- No information on item non-response was provided.

Guidance from the Validity Algorithm

Unclear if specifications consistent with evidence because evidence not presented (Box 1) → Information regarding threats to validity either not assessed or information not provided (Box 2) → Insufficient NOTE also that it is not clear whether or not the required score-level validation was conducted (Box 6)

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

RATIONALE: Need clarification regarding whether or not score-level validation was conducted. If not, it must be done and results must be adequate. Also, more information regarding potential threats to validity must be provided.

Committee pre-evaluation comments

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

2a.1 & 2b.1 Specifications: Reliability-Specifications

Comments:

**Specifications are clear. One would assume that all persons receiving services in a mental health setting (non-primary care) would fill out the survey. This measure could be consistently administered in a behavioral health treatment setting.

** The developers did not provide sufficient information about related to the evidence for the 17 measures to adequately discuss.

**Data elements are specified. A calculation algorithm is provided (The score for each multi-item measure is the adjusted average of the responses to composite items. The score for single item measures is the adjusted average score to that item for all patients in a given unit (e.g., plan). Adjustments are made using the CAHPS Macro which estimates a regression model in which all the plans are "absorbed" or fixed effects and a linear regression model is used to estimate adjusted plan scores after adjusting for self-reported mental health status, self-reported general health status, alcohol/drug treatment, age, education, and race/ethnicity) but is not sufficient to describe how the measure results are calculated. It does, however, allude to a SAS “CAHPS Macro” that can be used to calculate measure results

**I want to note that in excluding those who received care only from a primary care setting only, excludes the majority in the commercial population taking psychotropic medications prescribed from their PCP. I am not aware of any research for how many use this care as their sole BH care.

2a.2 Reliability Testing

Comments:

** Testing was conducted on 16,000 plan members from three states. They represented members of managed care, HMOs, fee-for-service, and public assistance programs. Information on reliability was only provided for the score items, and no information was provided for the other ECHO items. Reliabilities for items provided were sufficient, with the exception of the providing information about treatment options. Overall reliability for all the items could not be determined with the provided information.

**Data used in testing included surveys from approximately 16,000 plan members from three states (New Jersey, Minnesota, and an unknown MBHO). These data represented responses from beneficiaries from fee-for-service, managed care, commercial health maintenance organizations, and public assistance health programs. Data element reliability for the 5 multi-item PRO-PMs was calculated by examining 1) the internal consistency using Cronbach's alpha and 2) the item-total correlation (the correlation between the item score and the multi-item score if that item is removed). No testing information for the single-item PRO-PMs was provided.

Using data from New Jersey, the developer reports testing for measure score reliability using one-way analysis of variance. The results provided indicate that this analysis was done for the items in the survey, rather than for the PRO-PM measure results.

2b.2 Validity Testing

Comments:

** Content and construct validity testing conducted. The use of the focus groups with the providers and consumers, and the cognitive testing with the service recipients is standard. The explanation of the testing on all 17 items was insufficient.

**Validity was assessed via content validity, cognitive testing, and construct validation. Content validity: The developer noted that topic areas not initially assessed (e.g., adequacy of insurance coverage) was identified by the focus groups.

Cognitive testing: The developer states that testing ensured "that questions are consistently understood and that questions are understood in the manner intended by questionnaire developers".

Construct validation: In general there was correlation between survey measures and overall ratings.

2b.3.-2b7. Testing (Related to Potential Threats to Validity)

Comments:

** Exclusions appear appropriate if the purpose of the survey is to only assess those receiving services from mental health providers exclusively. The developers did not provide enough detail regarding their risk adjustment to determine if it was appropriately developed and tested.

Missing data discussion was related to two studies' response rates.

** The developer notes that there are no exclusions; however, patients who received behavioral health services only in primary care settings are excluded from the measure.

The developer does not provide any information about how many patients are actually excluded from the measure. I had difficulty following how the SDS variables were determined and what the analysis results mean. It was not clear how this measure distinguished meaningful differences.

The developers identify several analyses they conducted for both SDS and non-SDS factors in their case-mix model-building approach. It does not appear that a comparison was made to determine effects of risk-adjustment with inclusion of SDS factors versus without inclusion of SDS factors

The developer indicates missing data due to non-response is addressed via guidance on how to increase the number of returned surveys. They also report response rates of ranging from 36% to 65%, depending on type of plan (commercial vs. Medicaid) and depending on mode of administration using data are based on fielding of the survey in the late 1990's. No information on item non-response was provided.

** I want to note that in excluding those who received care only from a primary care setting only, excludes the majority in the commercial population taking psychotropic medications prescribed from their PCP. I am not aware of any research regarding how many use this care as their sole BH care. Patients also excluded who did not answer questions appropriately?

Criterion 3. Feasibility

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

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

- Data for these 17 PRO-PMs are collected via a survey that can be administered via mail, telephone, or mixed.
- Currently the responses to the ECHO survey are not captured in an electronic system; however, the developers report an on-going study to investigate using patient portals to collect such data.
- The developer indicates that all surveys and related materials are available free of charge on the AHRQ website.
- Recent information on typical response rates is not provided.

Questions for the Committee:

- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**The use of electronic methods at the clinic to collect information would be important for widespread use. The 17 items do not appear to be burdensome.

** Data for these 17 PRO-PMs are collected via a survey that can be administered via mail, telephone, or mixed. Currently the responses to the ECHO survey are not captured in an electronic system; however, the developers report an on-going study to investigate using patient portals to collect such data. The developer indicates that all surveys and related materials are available free of charge on the AHRQ website. Recent information on typical response rates is not provided.

** Insufficient 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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details

- The developers of the measure are not aware of specific uses of the 17 PRO-PMs included in this submission. They note use of technical assistance resources and visits/downloads from the ECHO webpage and interpret these as indication of the use of the measures.

Improvement results: No data on performance reported.

Unexpected findings (positive or negative) during implementation: None reported.

Potential harms: None reported.

Vetting of the measure: No information reported.

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?*
- *How has the measure been vetted in real-world settings by those being measure or others?*

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

RATIONALE: Little information on use or improvement resulting from use is available to the developers.

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use:

Comments:

**The measure is currently not being reported publically. This measure likely competes with a number of measures already being used at locations across the United States, including the MHSIP, which is mandated by SAMHSA. The information in the survey can be used to guide quality improvement efforts at both the plan and agency levels.

** The measure is not being publically reported. The developer notes that" The CAHPS Consortium makes surveys, including the ECHO surveys, available to the public for use but does not systematically track use. We have anecdotal information that states and plans are using ECHO data to stimulate and monitor improvement, but we do not systematically compile such data."

** Insufficient information. May be a useful survey but, as noted in preliminary analysis, not structured a as current PRO-PM and the submission has provided insufficient information in many areas, including this one. Is there a approved PRO-PM that could be used as a model for the developers?

Criterion 5: Related and Competing Measures

Related or competing measures

- N/A

Harmonization

- N/A

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because score-level reliability testing was not conducted and there is no information regarding potential vetting of the PRO-PMs by health plans or others.

Pre-meeting public and member comments

- No comments received.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0008

Measure Title: Experience of Care and Health Outcomes (ECHO) Survey

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: 12/27/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☒ Health outcome: Click here to name the health outcome

☒ Patient-reported outcome (PRO): Experiences with behavioral health care and perceived behavioral health improvements

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.12 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.

Structure > Clinical and clinician processes > Patient experiences > Responses to survey questions about patient experiences and perceived improvements in behavioral health

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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

We (Cleary, 2016) recently reviewed studies of the association between measures of patient-centered care (not necessarily focused on behavioral health care) and found broad support for the positive association between measures of patient centered care and other indicators of care quality. We identified just one study out of nearly three dozen that reported a negative correlation between patient experiences and clinical care quality. The authors of that study found a positive association between high quality patient-centered care and mortality and they suggested that providers may have been providing inappropriate care to improve "patient satisfaction." However, reanalysis of the data used in that study found such an association only for mortality that was not amenable to care, suggesting that the association observed was probably due to providers' giving better patient-centered care to sicker patients, such as those near the end of life (Xu et al. 2015, Elliott et al. 2013).

Some of the studies we reviewed did not find a significant association between patients' care experiences and clinical processes or outcomes (Anhang Price, Elliott, Cleary, et al. 2014), but that is not surprising, because individual quality indicators may or may not reflect quality of care in other areas (Wilson et al. 2007). There are a variety of possible

reasons why measures of patient experiences are correlated with other quality measures. One is that they are causally related. For example, it might be that in hospitals where there is better communication, patients are more likely to take medications appropriately, follow discharge instructions, and be less likely to be readmitted. It may also be, however, that they are correlated because better run health care organizations are likely to have better outcomes in multiple domains.

Anhang Price, R., M.N. Elliott, A.M. Zaslavsky, R.D. Hays, W.G. Lehrman, L. Rybowski, Edgman-Levitan., and P.D. Cleary. 2014. "Examining the role of patient experience surveys in measuring health care quality." *Med Care Res Rev* 71 (5):522-54.

Cleary PD, Evolving Concepts of Patient-Centered Care and the Assessment of Patient Care Experiences; Optimism and Opposition. J Health Pol, Policy & Law, 2016, 41 (4): 675-696.

Elliott, M.N., A.M. Haviland, P.D. Cleary, A.M. Zaslavsky, D.O. Farley, D.J. Klein, C.A. Edwards, M.K. Beckett, N. Orr, and D. Saliba. 2013. "Care experiences of managed care Medicare enrollees near the end of life." *J Am Geriatrics Soc* 61 (3):407-412.

Wilson, .IB., B.E. Landon, P.V. Marsden, L.R. Hirschhorn, K. McInnes, L. Ding, and P.D. Cleary. 2007. "Correlations among quality measures in HIV Care in the United States: a cross-sectional study of care sites in 30 states." *Br Med J* 335 (7629):1085-91.

Xu, X., E. Buta, R.A. Price, M.N. Elliott, R.D. Hays, and P.D. Cleary. 2015. Methodological considerations when studying the association between patient-reported care experiences and mortality. *Health Serv Res*. doi:10.1111/1475-6773.12264.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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

N/A

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

1a.4 OTHER SOURCE OF EVIDENCE

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

N/A

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.

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_evidence_attachment_12-27-2016-636192980306805249.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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)*

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

Donabedian and others have suggested that evaluating the process of care is one of the most direct ways to assess quality of care (Cleary, 2016). More recently, it has been suggested that one of eight ways that federal agencies can remove barriers to the delivery of effective behavioral care is measurement at the client, provider, organization, and population levels (Karakus, Ghose, et al.2016). The ECHO Survey assesses patient experiences with behavioral health services in such areas as getting treatment quickly, communication with clinicians, and information about treatment options. In moving away from global satisfaction questions, toward reports about specific well-defined aspects of care, the ECHO Survey more directly assesses quality of care than measures of "satisfaction".

The quality of behavioral health care is of significant concern because mental illnesses and alcohol and substance abuse impose substantial burdens on patients, employers, and the health care system. Payors and regulators are interested in assessing quality by comparing the experiences of patients receiving care from different health care organizations. A fundamental goal of CAHPS instruments is to provide evidence that can be used to improve health care and this was a major reason for developing the ECHO measures:

Shaul and colleagues noted that surveys such as the ECHO can identify aspects of the plan and treatment that are improvement priorities. Use of these data is likely to extend beyond the behavioral health plan to consumers, purchasers, regulators and policymakers, particularly since NCQA is encouraging behavioral health plans to use a similar survey for accreditation purposes.

To facilitate the use of such measures to stimulate quality improvement efforts, the CAHPS team reviewed the literature on strategies for improving aspects of care assessed in CAHPS surveys and developed: "The CAHPS Ambulatory Care Improvement Guide: Practical Strategies for Improving Patient Experience", which can be seen at:

http://www.ahrq.gov/cahps/quality-improvement/improvement-guide/improvement-guide.html:

Cleary, P. D. (2016). Evolving Concepts of Patient-Centered Care and the Assessment of Patient Care Experiences: Optimism and Opposition. Journal of health politics, policy and law, 3620881.

Karakus, M., Ghose, S. S., Goldman, H. H., Moran, G., & Hogan, M. F. (2016). "Big Eight" Recommendations for Improving the Effectiveness of the US Behavioral Health Care System. Psychiatric Services, appi-ps.

Shaul JA, Eisen SV, Stringfellow VL, Clarridge BR, Hermann RC, Nelson D, Anderson E, Kubrin AI, Leff HS, Cleary PD. Use of consumer ratings for quality improvement in behavioral health insurance plans. Jt Comm J of Qual Imp; 2001; 27: 216-229.

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 (4b) under Usability and Use.*
The CAHPS consortium does not compile or monitor performance data from ECHO users. Thus, it has limited information related to performance scores on the ECHO instrument.

Below are selected published studies describing the development and psychometric characteristics of the CABHS instrument, which was developed prior to the ECHO, and the subsequent ECHO instrument.

Shaul JA, Eisen SV, Stringfellow VL, Clarridge BR, Hermann RC, Nelson D, Anderson E, Kubrin AI, Leff HS, Cleary PD. Use of consumer ratings for quality improvement in behavioral health insurance plans. Jt Comm J of Qual Imp; 2001; 27: 216-229.

In 1998 and 1999, data were collected from five groups of adult patients in commercial health plans and five groups of adult patients in public assistance health plans with services received through four organizations, one of three MBHOs or a health system. Patients who received behavioral health care services during the previous year were mailed the CABHS survey. Non-respondents were contacted by telephone to complete the survey.
Response rates ranged from 49% to 65% for commercial patient groups and from 36% to 51% for public assistance patients. Getting treatment promptly from clinicians and aspects of care most influenced by health plan policies and operations, such as access to treatment and plan administrative services, received the least positive responses, whereas questions about communication received the most positive responses. In addition, questions about access and plan related aspects of quality showed the most inter-plan variability. Three of the organizations in this study focused quality improvement efforts on access to treatment.

Eisen, S.V., B. Clarridge, V. Stringfellow, J.A. Shaul, and P.D. Cleary. "Toward a National Report Card: Measuring Consumer Experiences with Behavioral Health Services. In B. Dickey and L. Sederer (Eds.), Improving Mental Health Care: Commitment to Quality. Washington, DC: APA Press, 2001. Chapter 9;115-134.

The CABHS survey was field tested with two groups of mental health consumers: commercially insured individuals (N=200), and Medicaid enrollees (N=300). Both groups were members of an HMO for which the behavioral health component was managed by an external managed behavioral health care organization. The survey was conducted by a survey research center that was independent of the HMO and the managed care organization.

Survey response rates reached an acceptable level (63%) for commercially insured consumers who were reachable by telephone. Response rates to mailed surveys alone were about 32%, highlighting the importance of telephone follow-up. A large percentage of consumers, particularly Medicaid enrollees, were not reachable by phone (55%). Of those reached by telephone, 57% participated in the survey.

The majority of respondents were female (80%), between the ages of 25 and 44 (61%) and were high school graduates (82%). Forty-nine percent were commercially insured; 51% were Medicaid beneficiaries. Almost two-thirds (64%) were Caucasian, 19% were Black or African-American, and 17% were members of other racial groups. Overall health was reported to be "fair" or "poor" by 27% of the sample. Mental health was reported to be "fair or "poor" by 34% of the sample. Respondents did not differ significantly from non-respondents in terms of age or sex.

Questions regarding behavioral health and substance abuse service use indicated that 86% reported receiving treatment for mental illness, personal or family problems and 76% of them reported taking prescription medications as part of their treatment.

Nine percent of the sample reported receiving services for alcohol abuse and 7% reported receiving services for drug abuse. Ten percent of respondents reported having received inpatient care.

Overall, responses to survey questions evaluating the care received suggested relatively positive experiences. Clinician-consumer interaction was rated more favorably than access to care. The highest rated aspects of care were items indicating clinician and office staff respect for consumers, frequency with which consumers felt listened to by clinicians, and clinicians' explanation of things in ways that consumers could understand. Least favorable ratings concerned accessibility to help in the evenings and on weekends.

Among questions about administrative burden and global evaluation of the insurance plan, responses were most favorable to the question about paperwork; this result was expected since all participants in this field test were HMO members for whom no paperwork was required. Respondents gave the least favorable responses to a question about handling of phone calls to the plan without a long wait. Overall evaluation of the plan (8.33 on a 10-point scale where 10 is the most positive rating) was almost identical to the overall evaluation of services (8.24).

One of the goals of the CABHS survey was to compare consumer experiences with different health plans. The field test was able to identify perceived differences between the two plans assessed (i.e., commercial plan and Medicaid). Consumers enrolled in the commercial plan rated timeliness of help on weekdays, evenings and weekends more highly than did those in the Medicaid plan. In addition, commercial plan members more often reported that they were told they could refuse treatment they did not want (87%) compared to Medicaid plan members (67%). On the other hand, Medicaid enrollees rated their health plan more highly overall than did consumers in the commercial plan (mean global rating of plan=8.80 for the Medicaid plan and 7.84 for the commercial plan, where 10 is the highest rating). Reported differences between the health plans were not associated with differences between the groups in demographic variables or health status of enrollees (Eisen et al. 1999).

Eisen SV, Shaul JA, Leff HS, Stringfellow V, Clarridge BR, Cleary PD. Toward a national consumer survey: Evaluation of the CABHS and MHSIP Instruments. J Behav Health Serv & Res; 2001; 28(3): 347-369.

This paper describes a study evaluating the Consumer Assessment of Behavioral Health Survey (CABHS) and the Mental Health Statistics Improvement Program (MHSIP) surveys. The purpose of the study was to provide data that could be used to develop recommendations for an improved instrument. Subjects were 3,443 adults in six behavioral health plans. The surveys did not differ significantly in response rate or consumer burden. Both surveys reliably assessed access to treatment and aspects of appropriateness and quality. The CABHS survey also reliably assessed features of the insurance plan, and the MHSIP survey reliably assessed treatment outcome. Analyses of comparable items suggested which survey items had greater validity. Results are discussed in terms of consistency with earlier research with these and other consumer surveys. Implications and recommendations for survey development, quality improvement and national policy initiatives to evaluate health plan performance are presented.

We do not have data over time for the ECHO but oher student using HCAHPS indicate that other CAHPS surveys can detect important temporal trends:

Elliott MN, Lehrman WG, Goldstein EH, Giordano LA, Beckett MK, Cohea CW, Cleary PD. Hospital survey shows improvements in patient experience. Health Aff, 2010; 29(11): 2061-2067. PMID: 21041749.

The comparison of scores on the HCAHPS survey for hospitals that reported data in 2008 and 2009 shows that after only one year of public reporting, hospitals that participated in the first public reporting experienced modest but meaningful improvements on all measures except for doctor communication, with the biggest gains in discharge information, hospital quietness, and staff responsiveness

Elliott MN, Cohea CW, Lehrman WG, Goldstein E, Cleary PD, Giordano LA, Beckett MK, Zaslavsky AM. Accelerating improvement and narrowing gaps: Trends in patients' experiences with hospital care reflected in HCAHPS public reporting. Health Serv Res, 2015, 50 (6): 1850-67.

HCAHPS scores increased by 2.8 percentage points from 2008 to 2011 in the most positive response category. Among the middle 95 percent of hospitals, changes ranged from a 5.1 percent decrease to a 10.2 percent gain overall. The greatest improvement was in for-profit and larger (200 or more beds) hospitals. Five years after HCAHPS public reporting began, meaningful improvement of patients´ hospital care experiences continues, especially among initially low-scoring hospitals, reducing some gaps among hospitals.

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 Epidemiologic Catchment Area (ECA) study and the National Comorbidity Survey (NCS) indicate that approximately 30% of the adult US general population ages 15-54 met diagnostic criteria for at least one mental disorder in the past 12 months (Regier, Kaelber et al. 1998). These surveys also found low rates of treatment of mental illness, for example the NCS found that only 13.3% of persons with a psychiatric problem used outpatient health care services for that problem in the past 12 months (Kessler, Zhao et al. 1999).

Few validated measures of mental health care are available for use (Shield, Campbell et al. 2003). Such measures are needed to indicate where variations in care exist and to help improve care where quality is found to be lacking (Seddon, Marshall et al. 2001). Researchers have shown that there are gaps between existing measures for mental health quality of care and the needs of consumers, providers, and policy makers (Hermann, Leff et al. 2000). A recent review of mental health quality measures found that only 12% have been assessed for reliability and 3% for validity (Hermann, Leff et al. 2000).

Variation in quality:

Studies have shown considerable geographic variation in quality of mental health services (Shield, Campbell et al. 2003). Another study suggest plan coverage is a source of variation in quality noting that compared to those without depressive symptoms the disadvantage was larger in Medicare Advantage than in Fee-For-Service for those with depressive symptoms (Martino, Elliott et al. 2016). In addition, studies using the ECHO Survey point to differences in mental health care quality as perceived by patients. In one study using the ECHO Survey, of 47 ECHO Survey questions tested, analysis of variance found statistically significant differences between the 12 health plans studied for all but 8 items (Shaul, Eisen et al. 2001).

Significant opportunity for improvement:

The variation in quality and population differences described in section 1b.2 and in the articles cited below, indicate that there are significant opportunities to reduce disparities in access to services and treatment by geography, race, and age.

Abel G, Mavaddat N, Elliott MN, et al. Primary care experience of people with longstanding psychological problems: evidence from a national UK survey. Int Rev Psychiatry 2011 Jan;23(1): 9-Feb.

Bell, C. C. and H. Mehta (1980). "The misdiagnosis of black patients with manic depressive illness." J Natl Med Assoc 72(2): 141-5.

Blazer, D. G., C. F. Hybels, et al. (2000). "Marked differences in antidepressant use by race in an elderly community sample: 1986-1996." Am J Psychiatry 157(7): 1089-94.

Borowsky, S. J., L. V. Rubenstein, et al. (2000). "Who is at risk of nondetection of mental health problems in primary care?" J Gen Intern Med 15(6): 381-8.

Conner, K. O., Copeland, V. C., Grote, N. K., Koeske, G., Rosen, D., Reynolds, C. F., & Brown, C. (2010). Mental health treatment seeking among older adults with depression: the impact of stigma and race. The American Journal of Geriatric Psychiatry, 18(6), 531-543.

Cook, B. L., Zuvekas, S. H., Carson, N., Wayne, G. F., Vesper, A., & McGuire, T. G. (2014). Assessing racial/ethnic disparities in treatment across episodes of mental health care. Health services research, 49(1), 206-229.

German, P. S., S. Shapiro, et al. (1985). "Mental health of the elderly: use of health and mental health services." J Am Geriatr Soc 33(4): 246-52.

Hall, L. L. and L. M. Flynn (1997). "NAMI´S managed care report card. National Alliance for the Mentally Ill." Eval Rev 21(3): 352-6.

Hermann, R. C., H. S. Leff, et al. (2000). "Quality measures for mental health care: results from a national inventory." Med Care Res Rev 57 Suppl 2: 136-54.

Kales, H. C., F. C. Blow, et al. (2000). "Race, psychiatric diagnosis, and mental health care utilization in older patients." Am J Geriatr Psychiatry 8(4): 301-9.

Kessler, R. C., S. Zhao, et al. (1999). "Past-year use of outpatient services for psychiatric problems in the National Comorbidity Survey." Am J Psychiatry 156(1): 115-23.

Martino SC, Elliott MN, Haviland AM, et al. Comparing the Health Care Experiences of Medicare Beneficiaries with and without Depressive Symptoms in Medicare Managed Care versus Fee-for-Service. Health Serv Res 2016 June;51(3): 1002-20.

Padgett, D. K., C. Patrick, et al. (1994). "Ethnicity and the use of outpatient mental health services in a national insured population." Am J Public Health 84(2): 222-6.

Regier, D. A., C. T. Kaelber, et al. (1998). "Limitations of diagnostic criteria and assessment instruments for mental disorders. Implications for research and policy." Arch Gen Psychiatry 55(2): 109-15.

Ross, E. C. (1997). "Managed behavioral health care premises, accountable systems of care, and AMBHA´S perms. American Managed Behavioral Healthcare Association." Eval Rev 21(3): 318-21.

Seddon, M. E., M. N. Marshall, et al. (2001). "Systematic review of studies of quality of clinical care in general practice in the UK, Australia and New Zealand." Qual Health Care 10(3): 152-8.

Shaul, J. A., S. V. Eisen, et al. (2001). Experiences of Care and Health Outcomes (ECHO) Survey Field Test Report: Survey Evaluation (unpublished report). Boston, MA.

Shield, T., S. Campbell, et al. (2003). "Quality indicators for primary care mental health services." Qual Saf Health Care 12(2): 100-6.

Sirey, J. A., B. S. Meyers, et al. (1999). "Predictors of antidepressant prescription and early use among depressed outpatients." Am J Psychiatry 156(5): 690-6.

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 (4b) under Usability and Use.*

The CAHPS Consortium does not collect or compile data from CAHPS surveys, including the ECHO, so it does not have direct data on disparities.

Leff and colleagues, however, used ECHO survey data to examine the relationship between cultural factors and quality of care.

Measurement of patient satisfaction is now considered essential for providing patient centered care and is an important tool for addressing health care disparities. However, little is known about how ethnically and racially diverse (ERD) groups differ in how they perceive quality, and widely used instruments for measuring perceived quality give little attention to cultural elements of care. Leff and colleagues, however, examined the relationship between the culturally determined beliefs and expectations of four ERD groups (African Americans, Latinos, Portuguese-speakers, and Haitians, total N = 160) and the technical quality of treatment for depression provided in four "culturally-specific" primary care clinics. Using data from the Experiences of Care and Health

Outcomes survey, chart reviews and focus groups, the study addressed a set of questions related to the psychometric properties of perceived care measures and the technical quality of care. The groups differed in preferred cultural elements except all preferred inclusion of religion. They did not differ in overall perceived quality. Technical quality was higher for Portuguese and Haitians than for African Americans and Latinos.

Martino et al., in a study of adults who received behavioral health services and who completed the ECHO Survey, found that commercial insurance coverage was associated with better general and mental health, on average, than Medicaid insured (3.44 versus 2.80 for general health, P<.001, and 3.35 versus 2.89 for mental health, P<.001).

Leff, H. S., Chow, C., Wieman, D. A., Ostrow, L., Cortés, D. E., & Harris, T. (2016). Measurement of Perceived and Technical Quality of Care for Depression in Racially and Ethnically Diverse Groups. Journal of Immigrant and Minority Health, 1-9.

Martino SC, Elliott MN, Haviland AM, et al. Comparing the Health Care Experiences of Medicare Beneficiaries with and without Depressive Symptoms in Medicare Managed Care versus Fee-for-Service. Health Serv Res 2016 June;51(3): 1002-20.

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**

Quality of mental health services differs by a variety of demographic characteristics. In the 2009 English GP Patient Survey, 5.7% of 2,163,456 respondents reported that they had a long-standing psychological or emotional condition. In an unadjusted regression model, respondents with long-standing emotional or psychological conditions rated their experiences worse than people without such problems, with scores which were up to 3 percentage points lower on individual survey items. However, after controlling for age, gender, ethnicity, deprivation and self-reported general health, people with long-standing psychological or emotional problems had slightly higher scores on 16 out of the 18 survey items, though with the equivalent of less than 2 percentage points difference for most items (Abel, Mavaddat, et al. 2011).

Regarding the diagnosis and treatment of mental illness race appears to be an important determinant. Doctors are less likely to identify blacks as depressed (Bell and Mehta 1980; Borowsky, Rubenstein et al. 2000) or to prescribe anti-depressant medications for blacks, compared to whites (Sirey, Meyers et al. 1999; Blazer, Hybels et al. 2000). Also, in general, blacks are less likely to use mental health services compared to whites, even when similarly insured (Padgett, Patrick et al. 1994; Kales, Blow et al. 2000).) and blacks and Latinos had shorter episodes of care and fewer psychotropic drug fills (Cook, Zuvekas, et al. 2014).

In addition, older adults are less likely to be treated for mental illness from specialty providers compared to younger adults (German, Shapiro et al. 1985) and depressed older adults perceived a high level of public stigma and were not likely to be currently engaged in or seek mental health treatment (Conner, Copeland, et al. 2010).

(see 1b.3 for references)

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.ahrq.gov/cahps/surveys-guidance/echo/index.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)*
No data dictionary **Attachment:**

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.
The ECHO has been updated with minor wording changes so that it is consistent with the family of CAHPS surveys. As part of this process, several focus groups were conducted and some terminology was updated to be more consistent with widely understood words and phrases.

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).
No changes form original specification: The ECHO survey measures patient-centered care by asking about patient experiences with behavioral health care (mental health and substance abuse treatment) and the organizations that provide or manage the person´s treatment and health outcomes.

The survey and instructions are available at:
www.qualityforum.org/pdf/ambulatory/txECHOALL(onepager&specs&survey)03-23-07.pdf
Measure developer/instrument web site:
www.cahps.ahrq.gov/content/products/ECHO/PROD_ECHO_MBHO.asp?p=1021&s=214
The composite measures' component items can be found on the document titled "Reporting Measures for the ECHO Survey 3.0" (Document No. 209 – 8/31/06) available for download at http://www.ahrq.gov/cahps/surveys-guidance/echo/instructions/index.htmlNo changes form original specification: The ECHO survey measures patient-centered care by asking about patient experiences with behavioral health care (mental health and substance abuse treatment) and the organizations that provide or manage the person´s treatment and health outcomes.

The survey and instructions are available at:
www.qualityforum.org/pdf/ambulatory/txECHOALL(onepager&specs&survey)03-23-07.pdf
Measure developer/instrument web site:
www.cahps.ahrq.gov/content/products/ECHO/PROD_ECHO_MBHO.asp?p=1021&s=214

The composite measures' component items can be found on the document titled "Reporting Measures for the ECHO Survey 3.0" (Document No. 209 – 8/31/06) available for download at http://www.ahrq.gov/cahps/surveys-guidance/echo/instructions/index.html

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).
No changes: Responses from all survey respondents, or for selected items, all respondents who respond appropriately to screening questions.

Eligible respondents are health plan or MBHO patients who have been continuously reenrolled for the past 12 months, 18 years or older, with diagnostic or procedural code in administrative records.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All survey respondents, or for selected items, all respondents who respond appropriately to screening questions.

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).
No changes: The denominator is the number of survey respondents in a health plan or managed behavioral healthcare organization (MBHO).

Eligible respondents are health plan or MBHO patients who have been continuously reenrolled for the past 12 months, 18 years or older, with diagnostic or procedural code in administrative records.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
No changes: Patients who received behavioral health services only in primary care settings (e.g. psychotropic medications from their primary care physician) are not included.

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.)*
Patients who received behavioral health services only in primary care settings (e.g. psychotropic medications from their primary care physician) in the preceding 12 months are not included.

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)
Statistical risk model
If other:

S.12. **Type of score:**
Continuous variable, e.g. average
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
The score for each multi-item measure is the adjusted average of the responses to composite items. The score for single item measures is the adjusted average score to that item for all patients in a given unit (e.g., plan). Adjustments are made using the CAHPS Macro which estimates a regression model in which all the plans are "absorbed" or fixed effects and a linear regression model is used to estimate adjusted plan scores after adjusting for self-reported mental health status, self-reported general health status, alcohol/drug treatment, age, education, and race/ethnicity.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
The sample should be a random sample of eligible patients and needs to be large enough to yield 411 completed surveys per health care organization, a cost-effective method shown to produce statistically valid survey comparisons.

Proxy respondents are not allowed.

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.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Guidance provided on the web site page specified in 5.1

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

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Available at: http://www.ahrq.gov/cahps/surveys-guidance/echo/instructions/index.html

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

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)*
Behavioral Health : Outpatient
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
NQF_testing_attachment_12-27-2016.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. (Do not remove prior testing information – include date of new information in red.)
Yes

2.2 <u>For maintenance of endorsement</u>
Has additional empirical validity testing of the measure score been conducted? If yes, please provide results in the Testing attachment. (Do not remove prior testing information – include date of new information in red.)
No

2.3 <u>For maintenance of endorsement</u>
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Yes - Updated information required during the SDS Trial Period is included

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

Measure Number (*if previously endorsed*)**:** 008
Measure Title: Experience of Care and Health Outcomes (ECHO) Survey
Date of Submission: 12/26/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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 data described below come predominately from data from two studies that have been described int eh publications below:

Shaul JA, Eisen SV, Stringfellow VL, Clarridge BR, Hermann RC, Nelson D, Anderson E, Kubrin AI, Leff HS, Cleary PD. Use of consumer ratings for quality improvement in behavioral health insurance plans. Jt Comm J of Qual Imp; 2001; 27: 216-229.

In 1998 and 1999, data were collected from five groups of adult patients in commercial health plans and five groups of adult patients in public assistance health plans with services received through four organizations, one of three MBHOs or a health system.

Eisen, S.V., B. Clarridge, V. Stringfellow, J.A. Shaul, and P.D. Cleary. "Toward a National Report Card: Measuring Consumer Experiences with Behavioral Health Services. In B. Dickey and L. Sederer (Eds.), Improving Mental Health Care: Commitment to Quality. Washington, DC: APA Press, 2001. Chapter 9;115-134.

The CABHS survey was field tested with two groups of mental health consumers: commercially insured individuals (N=200), and Medicaid enrollees (N=300). Both groups were members of an HMO for which the behavioral health component was managed by an external managed behavioral health care organization.

1.3. What are the dates of the data used in testing? 2001-2015 (only earlier data published and described here)

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.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <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*)

The performance of the ECHO measures has been tested in multiple settings. Below are selected published studies describing the units and patients included in testing during development and psychometric characteristics of the CABHS instrument, which was developed prior to the ECHO, and the eventual ECHO instrument.

Shaul JA, Eisen SV, Stringfellow VL, Clarridge BR, Hermann RC, Nelson D, Anderson E, Kubrin AI, Leff HS, Cleary PD. Use of consumer ratings for quality improvement in behavioral health insurance plans. Jt Comm J of Qual Imp; 2001; 27: 216-229.

In 1998 and 1999, data were collected from five groups of adult patients in commercial health plans and five groups of adult patients in public assistance health plans with services received through four organizations, one of three MBHOs or a health system. Patients who received behavioral health care services during the previous year were mailed the CABHS survey. Non-respondents were contacted by telephone to complete the survey. Response rates ranged from 49% to 65% for commercial patient groups and from 36% to 51% for public assistance patients.

Eisen, S.V., B. Clarridge, V. Stringfellow, J.A. Shaul, and P.D. Cleary. "Toward a National Report Card: Measuring Consumer Experiences with Behavioral Health Services. In B. Dickey and L. Sederer (Eds.), Improving Mental Health Care: Commitment to Quality. Washington, DC: APA Press, 2001. Chapter 9;115-134.

The CABHS survey was field tested with two groups of mental health consumers: commercially insured individuals (N=200), and Medicaid enrollees (N=300). Both groups were members of an HMO for which the behavioral health component was managed by an external managed behavioral health care organization. The survey was conducted by a survey research center that was independent of the HMO and the managed care organization.

The majority of respondents were female (80%), between the ages of 25 and 44 (61%) and were high school graduates (82%). Forty-nine percent were commercially insured; 51% were Medicaid beneficiaries. Almost two-thirds (64%) were Caucasian, 19% were Black or African-American, and 17% were members of other racial groups. Overall health was reported to be "fair" or "poor" by 27% of the sample. Mental health was reported to be "fair or "poor" by 34% of the sample. Respondents did not differ significantly from non-respondents in terms of age or sex.

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*)

See response to 1.5

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.

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

The ECHO survey asks about several patient characteristics, including self-reported mental health status, general health status, age, gender, race, ethnicity, education, income, and whether care was being sought for alcohol or drug use.

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*)

The psychometric properties of the ECHO Survey measures have been thoroughly tested. The results described below come from testing in 3 states, in fee-for-service and managed care plans, commercial health maintenance organizations, and public assistance health programs. Patients who were 18 years of age or older, enrolled in the health plan when the samples were drawn, and received behavioral health counseling or treatment services in the 12 months before the survey was administered were eligible for the study. Eligible treatment services included outpatient or ambulatory treatment sessions and services such as partial or day treatment services for mental illness, personal or family problems, and alcohol or drug dependency. Eligible services were identified according to diagnostic and procedural criteria established by the National Committee for Quality Assurance (NCQA) to compute inpatient and outpatient chemical dependency and mental health utilization rates. A total of approximately 16,000 plan members were surveyed.

New Jersey and Minnesota sampled a percentage of individuals who had not received behavioral health care services to maintain the confidentiality of the information. Individuals who did not receive services were excluded from these analyses. Only one individual was surveyed from each household. New Jersey and MBHO samples were also limited to those who received services from a specialty behavioral health clinician and who were continuously enrolled in their health plan for the 12 months preceding the survey.

Disabled FFS enrollees were underrepresented in the FFS sample and the sample was not representative of the FFS population. Each of the seven New Jersey health plans was asked to select a probability sample from their commercial enrollees. A few plans did not have an adequate number of patients receiving services; in these cases, all eligible members were selected (n=7,124; range: 664 per plan to 1,149 per plan). Probability samples of commercial and Medicaid enrollees in three health plans were selected by the MBHO (n=2,100, 1,500 commercial, 600 Medicaid). Item non-response rates, applicability rates, and the percent of respondents choosing the most positive or negative response option were computed for each item.

Reliability of the composite measures was estimated using Cronbach's alpha coefficient, a measure of internal consistency. Item-to-total correlations and Cronbach's alpha if an item was removed from the composite were also computed for the items in each scale. Because the purpose of the survey is to facilitate comparisons among health care organizations, we also used a one-way analysis of variance to compute plan level reliability estimates in New Jersey with plans as the between groups factor ($MS_{between}$-MS_{within})/$MS_{between}$). Since the Minnesota samples were designed to collect information about managed care and fee-for-service health plans overall, samples were not adequate to compute this statistic.

Redundancy of the measures was ascertained by computing correlations among them. The relationship between survey and the overall treatment rating and plan ratings was ascertained by correlating each measure with the ratings. To estimate crite validity of the overall plan rating, we estimated correlations between the plan rating and reports about intent to switch health pl New Jersey and Minnesota. In Minnesota, we also calculated correlations between the plan rating and reports about willingness recommend the health plan to friends and family members, level of confidence the plan would continue to meet the member's fa

needs, and level of satisfaction**.** Preliminary analyses revealed that results were similar for the FFS disabled and non-disabled gro the FFS and managed care patients in Minnesota. Results were also similar for the MBHO's commercial and Medicaid patients. T estimates based on pooled samples are reported for each site.

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*)

Five summary measures (composites) were created and tested for their psychometric properties: getting treatment quickly, how well clinicians communicate, perceived improvement, getting treatment information from the plan, and information about treatment options. Reliability as estimated by Cronbach's alpha (internal consistency reliability) was 0.72 or higher for 4 of the 5 composites across sites. The reliability for the *information about treatment options* composite was 0.65. Item-total correlations were also assessed for the 31 single items with their respective composites. All but four of these single items had item total correlations above 0.50 across sites.

Scale Reliability Statistics

		Minnesota		**MBHO**		**New Jersey**	
Reporting Measure	**Description**	**Coefficient alpha**	**Item-to-total correlation**	**Coefficient alpha**	**Item-to-total correlation**	**Coefficient alpha**	**Item-to-total correlation**
Getting treatment quickly	Get urgent treatment as soon as needed	0.80 (n=600)	0.71	0.72 (n=162)	0.59	0.84 (n=458)	0.78
	Get appointment as soon as wanted		0.71		0.57		0.75
	Get help by telephone		0.55		0.48		0.58
How well clinicians communicate	Clinicians listen carefully	0.86 (n=1766)	0.71	0.85 (n=829)	0.70	0.89 (n=2154)	0.78
	Clinicians explain things		0.65		0.68		0.73
	Clinicians show respect		0.74		0.69		0.79
	Clinicians spend enough time		0.65		0.64		0.72
	Feel safe with clinicians		0.57		0.54		NA
	Involved as much as you wanted in treatment		0.56		0.59		0.67
Perceived Improvement	Compare ability to deal with daily problems to 1 year ago	0.83 (n=1769)	0.68	0.89 (n=837)	0.77	0.89 (n=2102)	0.77

Reporting Measure	Description	Coefficient alpha	Item-to-total correlation	Coefficient alpha	Item-to-total correlation	Coefficient alpha	Item-to-total correlation
	Compare ability to deal with social situations to 1 year ago		0.69		0.75		0.73
	Compare ability to accomplish things to 1 year ago		0.70		0.78		0.79
	Compare ability to deal with symptoms or problems to 1 year ago		NA		0.75		0.77
Getting treatment and information from the MBHO or Plan	Getting clinician happy with	0.80 (n=68)	0.41	0.75 (n=294)		0.84 (n=136)	0.62
	Delays in treatment while wait for plan approval		0.75		0.60		0.58
	Problem getting necessary treatment		0.41				0.66
	Understanding information about treatment in booklets or on the web		0.62				0.63
	Helpfulness of customer service		0.62		0.60		0.65
	Filling out paperwork		0.58				0.55
Informed about treatment options	Told about self-help or consumer run programs	0.66 (n=1766)	0.49	0.65 (n=818)	0.48	NA[a]	NA
	Told about different treatments that are available for condition		0.49		0.48		NA

=not applicable

New Jersey did not ask the item about different types of treatment

Plan level reliability estimates were above 0.7 for seven items, were between 0.4 and 0.7 for four items, and were between 0 and .4 for 2 items. For three items the reliability estimates were less than zero because there was insufficient variability among plans. See Table at end of this section for more detail.

The overall behavioral health care plan rating had the most inter-plan variability for both commercial and public assistance behavioral health plans (Table 5), although plan membership explained only about 3% of the variation in commercial plan ratings and 6% of the variation in public assistance plan ratings. Overall treatment and main clinician ratings did not vary significantly among the commercial plans. Public assistance plans demonstrated significant variation for these overall ratings, although the ratio of between plan variation to within plan variation was less than for the overall plan rating.

Items about the plan tended to demonstrate more variability among both commercial and public assistance plans than items about communication with clinicians, office staff, information provided by clinicians, and getting treatment quickly. Five of six items about the health plan discriminated among commercial plans and three of six items discriminated among the public assistance plans. In comparison, two of 11 items about communications with clinicians, office staff, and information provided by clinicians discriminated among commercial plans and three of the 11 items discriminated among public assistance plans. Two of the four items about getting treatment quickly discriminated only among the commercial plans.

Items about the plan, including delays in treatment while waiting for plan approval and getting necessary treatment showed substantial variation among commercial plans. Filling out paperwork, understanding information in written materials, and getting help also differentiated among the commercial plans. Questions about treatment delays, written materials, and customer service, varied significantly among public assistance plans.

The amount the respondent was helped by treatment demonstrated the largest variation among public assistance plans for items about counseling or treatment. Reports about this item for commercial plans did not vary significantly. Items about getting an appointment as soon as wanted, waiting in the office for 15 minutes or more to see the clinician, involvement in treatment decisions, and office staff also differentiated among commercial plans. Questions about clinicians who listened carefully and being told about the right to refuse treatment discriminated among public assistance plans.

Variation of consumers' evaluations among commercial and among public assistance behavioral care plans.

Item Number	Items	Commercial Differentiation Among Plans (F-statistic) (n=5)		Public Assistance Differentiation Among Plans (F-statistic) (n=5)
	Global Ratings			
23.	Rating of all treatment or counseling	0.80		4.07**
25.	Rating of clinician who provides most of your treatment or counseling	0.30		3.59**
40.	Rating of health plan	11.46***		10.35***
	Getting Treatment Quickly			

Item Number	Items	Commercial Differentiation Among Plans (F-statistic) (n=5)		Public Assistance Differentiation Among Plans (F-statistic) (n=5)
	People's experiences in getting treatment or counseling quickly:			
3.	When you needed to see a clinician right away, how often did you get the treatment or counseling as soon as wanted?	0.17		0.62
5.	How often did you get an appointment for treatment or counseling as soon as you wanted?	3.21*		0.68
7.	How often did you get the help or advice you needed by telephone?	1.09		0.98
	People's experiences waiting for treatment or counseling at their clinicians' offices:			
11.	How often did you wait in the office or clinic more than 15 minutes past your appointment time? (item was reverse coded)	3.31*		1.09
	How Well Clinicians Communicate			
12.	How often did clinicians listen carefully to you?	0.36		2.44*
13.	How often did clinicians explain things in a way you could understand?	0.86		0.72
14.	How often did clinicians show respect for what you had to say?	0.81		2.04
15.	How often did clinicians spend enough time with you?	1.66		2.35
18.	How often did clinicians give you reassurance and support?	0.88		0.83
19.	How often were you involved as much as you wanted in decisions about your treatment?	2.39*		1.02
	Courteous and Helpful Office Staff			
10.	How often did office staff treat you with courtesy and respect?	3.90**		0.85
	Information Giving by Clinicians			
17.	Did your clinician tell you about the risks and benefits of medications you have taken?	1.28		1.90
21.	Did clinicians tell you about different kinds of treatment available for your condition?	1.07		1.70
22.	Did your clinicians tell you that you have the right to refuse treatment that you do not want?	0.58		3.02*
	Amount Helped by Treatment			
21.	How much were you helped by the treatment or counseling you received?	1.48		5.13***
	The Health Plan			
	Getting Needed Treatment or Counseling:			
32.	With the choices your health plan gave, how much of a problem was it to get a clinician you are happy with?	2.37		2.23
34.	How much of a problem were delays in treatment while waiting for approval from your health plan?	11.20***		2.53*

Item Number	Items	Commercial Differentiation Among Plans (F-statistic) (n=5)		Public Assistance Differentiation Among Plans (F-statistic) (n=5)
37.	How much of a problem was it to get treatment you or a clinician believed necessary?	7.38***		2.29
	Plan Administrative Services:			
36.	How much of a problem did you have with paperwork for your health plan?	2.42*		0.98
39.	How much of a problem was it to find or understand information in the written materials?	4.98***		6.83***
41.	How much of a problem was it to get help when you called your health plan's customer service?	5.12***		4.16**

Stars indicate significant differences among plans after controlling for age, gender, cost-sharing, mental health status and education. *p<.05, ** p<.01, *** p<.001

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 results provide evidence that ECHO based measures are generally reliable. The one area that appears to assess reliably is providing information about treatment options. This measure was kept in the instrument because of the salience of providing accurate information and shared decision making to both users and providers of behavioral health services.

2b2. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Face Validity:

The ECHO Survey was developed as part of the CAHPS development process, and began with a study that compared the Consumer Assessment of Behavioral Health Services (CABHS) instrument to the Mental Health Statistics

Improvement Program's (MHSIP) consumer survey (Eisen, Shaul et al. 1999). The CABHS-MHSIP evaluation project included a survey of 3,400 adults enrolled in one of six behavioral health plans. Three focus groups were conducted with mental health service consumers with a broad range of experience with mental health services including for depression, anxiety, schizophrenia, and substance abuse. The focus groups helped identify content areas that had not been previously assessed, e.g. the adequacy of insurance coverage (Eisen, Shaul et al. 1999). Later cognitive testing was performed to ensure that questions are consistently understood and that questions are understood in the manner intended by questionnaire developers. Finally, some direct measures of outcomes are assessed, e.g. patient perceived improvement in the last 12 months, and current health status. Thus, all the information from CAHPS literature reviews, consumer input, focus groups, and cognitive interviews, was considered when developing the ECHO Survey. In addition, an advisory group, the ECHO Survey Development Team was created to make recommendations for the design and content of the survey. This group included behavioral health consumers, clinicians and behavioral health policy experts, including representatives from CAHPS, the National Committee for Quality Assurance (NCQA), NCQA's Behavioral Health Measurement Advisory Panel, the Mental Health Statistics Improvement Program, the Human Services Research Institute, the Center for Mental Health Services, the Washington Circle Group, the American Managed Behavioral Health Care Association, and the National Alliance for the Mentally Ill.

Eisen, S. V., J. A. Shaul, et al. (1999). "Development of a consumer survey for behavioral health services." Psychiatr Serv **50**(6): 793-8.

Empirical Validity Testing:

For a description of a study done to test the reliability and validity of the ECHO measures, see the description of methods in section 2a2.

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

Validity: Correlations between survey measures and overall ratings

(Minnesota (top), MBHO (middle), New Jersey (bottom).

All correlations are significant (p<.001) unless noted by *s or ns.

Measure:	**Treatment Rating**	**Plan Rating**
Getting treatment quickly	0.44	0.28
	0.37	0.20
	0.50	0.31
How well clinicians communicate	0.71	0.41
	0.72	0.20
	0.75	0.29
Perceived improvement	0.30	0.22
	0.32	0.16
	0.36	0.16
Getting treatment and information from the plan or MBHO	0.43	0.51
	0.17	0.52

	0.32	0.66
Informed about treatment options	0.17 0.43 0.11[a]	0.11 0.13 0.03(ns)[a]
Office wait	0.25 0.21 0.30	0.15 0.06 (ns) 0.10
Told about medication side effects	0.25 0.24 0.26	0.13 0.11* 0.11
Including family & friends	0.14 0.09* 0.12	0.06* 0.02 (ns) 0.02 (ns)
Information to manage condition	0.47 0.48 0.53	0.28 0.18 0.21
Informed about patient rights	0.17 0.21 0.22	0.11 0.15 0.14
Patient feels he or she could refuse treatment	0.17 0.19 0.16	0.14 0.00 (ns) 0.03 (ns)
Privacy	0.37 0.25 0.32	0.19 0.07* 0.16
Cultural competence	0.35 0.84 0.72	0.14* 0.24 (ns) 0.29
Amount helped	0.65 0.63 NA	0.31 0.21 NA

Treatment after benefits are used up	0.04 (ns)	0.23*
	0.15 (ns)	0.27
	0.05 (ns)	0.12*

A. a. New Jersey only asked the self-help group item for the information about treatment scale. The item about being told about different types of treatment was not asked. *p<0.05, **p<0.01, no stars=p<0.001, ns=p->0.05.

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

The ECHO based measures have face validity and generally perform well according to standard empirical criteria.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

_____

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

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

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

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

2b4.1.1 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.

Background

Increasingly, plan level summaries of patient experiences are being used in public reports about care quality (Goldstein et al. 2001) and by accreditation agencies (NCQA 2005). Policy makers and payors have become interested in assessing quality by comparing the experiences of patients across behavioral health organizations in an effort to increase accountability and better inform quality improvement efforts (Eisen et al. 2001a). However, comparisons may be misleading if scores are not adjusted for differences in the underlying characteristics of the patient populations across

plans. For example, patients' health status or age may influence how patients evaluate their experiences independent of the quality of care received. Moreover, quality of care may differ between certain types of patients (e.g., younger and older patients) in a consistent way within all health plans. If plans are compared using the average score of all enrollees, relative rankings may not accurately reflect quality of care differences. Ideally, health plan ratings would reflect how plans would be rated by an identical patient population. Thus, it is important to control for underlying patient characteristics when calculating plan-level scores. Even if the adjustment to scores is modest, it is important to adjust scores for differences in patient characteristics to ensure both the actual and perceived fairness of the comparisons.

To address these issues, we first identified patient characteristics that are associated with consumer ratings of behavioral health plans. Second, we developed a statistical model to adjust plan-level summaries of patients' ratings of behavioral health care plans so that ratings more accurately reflect quality of care received. Because Commercial and Medicaid plan populations may systematically differ, we develop separate models for these two plan types.

A priori, we expect self-reported mental health status, general health status, and age to be important adjusters based on prior research in the general health care sector (Zaslavsky et al. 2001). Education, income, gender, race/ethnicity, and seeking care for alcohol or drug use may also be important case-mix adjusters and are tested.

This study used data collected during the field test of the ECHO™ survey (Shaul et al. 2001), as well as survey data submitted voluntarily by health plans to the National CAHPS Benchmarking Database (NCBD 2005). These data were combined to achieve the largest sample of ECHO survey respondents available to date. The field test data included responses of enrollees in six Medicaid plans in Minnesota and seven commercial plans in New Jersey. The NCBD data included responses of enrollees in three Medicaid plans and five commercial plans in Colorado, Florida, New York, and Ohio.

Health plan enrollees were eligible to receive the survey if they were 18 years or older, had been continuously enrolled during the previous year, and had a diagnostic or procedural code in administrative records indicating they had received behavioral health services in the preceding year. Eligible services included treatment for mental illness, personal or family problems, and/or treatment for alcohol or drug use that was provided by a specialty behavioral health care provider in an outpatient, inpatient, or partial or day treatment setting. Enrollees who received behavioral health services only in primary care settings (e.g., enrollees who received psychotropic medications from their primary care physician) were not eligible. These criteria were consistent with those used by the National Committee on Quality Assurance (NCQA) to compute mental health and chemical dependency utilization rates (NCQA 2000).

Sampling procedures varied across the sites. For the field test data, Minnesota pooled eligible enrollees from six Medicaid plans and drew a random sample of 2,500 individuals. In New Jersey, four commercial plans selected simple random samples of 1,400 enrollees from their eligible population, while three plans with insufficient enrollment selected all eligible enrollees. The resulting samples ranged from 842 to 1,426. For the NCBD data, in New York, 300 survey recipients were sampled in each of two Medicaid plans, and 750 were sampled in each of two commercial plans. In Ohio, 300 survey recipients were sampled from the larger of two commercial plans, while 200 respondents were sampled from the smaller plan. Nearly 1,300 were sampled in a commercial plan in Florida, and 1,110 were sampled in a Medicaid plan in Colorado. In total, 14,482 behavioral health care patients were sampled. In all sites, only one plan enrollee was selected from each household. In all cases, the surveys were administered by independent survey vendors between September 2000 and February 2002. Half of the sites sent one or two mailings with telephone follow-up of non-respondents (at least six calls), while half relied solely on mailings.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**

The dependent variables used to develop the case mix adjustment models consisted of a global rating of all counseling or treatment received in the previous 12 months, a global rating of the health plan that managed that care and

22 questions that asked about specific experiences in four domains (timely access to care, patient-provider interaction, treatment information, and health plan approval and service). Specific questions are noted in an Appendix available on the journal website. To determine plan scores for the four domains, we calculated the mean of responses to the questions in each domain.

Eight variables were tested as potential case mix adjusters. These included self-reported general health status, self-reported mental health status, whether the respondent's reasons for receiving counseling or treatment in the past year included getting help for alcohol or drug use, age, gender, race/ethnicity, education and income. Income was available only for enrollees in seven commercial plans.

Because public reports and accreditation agencies use plan summaries to assess plan quality, we focus on the adjustment of such summary statistics. To be a relevant adjuster, a patient characteristic must be a significant predictor of scores and vary in distribution across plans. For example, if self-reported health status was not correlated with CAHPS scores, or the distribution of health status did not differ across health plans, it would not be an important case mix adjustor. Measures of these two criteria can be combined into a summary measure of "explanatory power" that can be used to compare the relative impact of potential adjusters (O'Malley et al. 2005; Zaslavsky et al. 2001). To determine predictive power, each individual-level rating was regressed on each adjuster in separate linear models. Dummy variables for plans were included in the predictive models, so the resulting coefficients were estimates of the within-plan effects. To measure how much the patient characteristic varied across plans, a variance ratio (the ratio of the adjuster's between-plan variance to its within-plan variance) was calculated. The explanatory power (EP) of each adjuster relative to each rating and report item was calculated by multiplying the adjuster's predictive power (contribution to R^2) by its variance ratio (Zaslavsky 1998). While predictive power refers to the contribution of the variable to predicting individual responses, explanatory power refers to the contribution of the variable to explaining differences among mean responses for groups (plans, in this case).

Adjusters that were significantly associated with patient reports and had high explanatory power were selected for the base model. Once the base model was determined, the remaining variables were retested by computing their incremental explanatory power after controlling for adjusters in the base model.

In the initial models, linear specifications of ordinal variables were included. This specification assumes that the difference between levels is constant (i.e., that the difference in ratings between poor and fair health is the same as the difference in ratings between very good and excellent health). Two advantages of this specification are that it results in a more parsimonious model, allowing for testing of interactions of case mix adjusters with plan once the base model is specified, and that it is consistent with other CAHPS case mix adjustment models. The disadvantage of this specification is it may lead to a suboptimal specification if the assumptions regarding the effect of differences in ratings are incorrect. To test the appropriateness of using linear specifications, we re-estimated our initial prediction regressions using categorical specifications of ordinal variables. We then compared the adjusted R-squared of the linear specification with the adjusted R-squared of the categorical specification to determine whether use of the linear specification was appropriate.

Once the final adjustment model(s) was specified, the CAHPS® analysis program (www.cahps.ahrq.gov) was used to calculate unadjusted and adjusted plan means for the two global ratings and four composite summary measures. Two plans with fewer than 30 respondents were excluded, leaving 19 plans in the sample for the impact analysis. The unadjusted and adjusted plan scores were centered to have a mean of zero so that the impact of alternative models could be compared. The impact of adjustment on plan scores was summarized by three measures: the mean absolute adjustment, the largest positive adjustment and the largest negative adjustment. In public reports, the relative ranking of plans may be important to consumers. The impact of adjustment models on plan rankings was summarized by two measures: Kendall tau correlations and the percentage of all possible pairs of plans whose rank-order changed post-adjustment. Finally, the uniformity of each adjuster's coefficient across plans was assessed by testing the interaction between the adjuster and a group (plan) variable in each model.

REFERENCES

Cleary, P. D. and S. Edgman-Levitan. 1997. "Health care quality. Incorporating consumer perspectives." *JAMA* 278(19):1608-12.

Cleary, P. D., S. Edgman-Levitan, W. McMullen, and T. L. Delbanco. 1992. "The relationship between reported problems and patient summary evaluations of hospital care." *Qual Rev Bull* 18(2):53-9.

Dow, M. G., T. L. Boaz, and D. Thornton. 2001. "Risk adjustment of Florida mental health outcomes data: concepts, methods, and results." *J Behav Health Serv Res* 28(3):258-72.

Eisen, S. V., B. Clarridge, V. Stringfellow, J. A. Shaul, and P. D. Cleary. 2001a. "Toward a national report card: Measuring consumer experiences with behavioral health services." In *Achieving quality in psychiatric and substance abuse practice: Concepts and case reports*, edited by B. Dickey and L. Sederer, Washington, DC: APA Press.

Fremont, A. M., P. D. Cleary, J. L. Hargraves, R. M. Rowe, N. B. Jacobson, and J. Z. Ayanian. 2001. "Patient-centred processes of care and long-term outcomes of myocardial infarction." *J Gen Int Med* 16:800-8.

Goldstein, E., P. D. Cleary, K. M. Langwell, A. M. Zaslavsky, and A. Heller. 2001. "Medicare Managed Care CAHPS: A Tool for Performance Improvement." *Health Care Finan Rev* 22(3):101-07.

Hall, J. A., M. A. Milburn, and A. M. Epstein. 1993. "A causal model of health status and satisfaction with medical care." *Med Care* 31(1):84-94.

Heflinger, C. A., C. G. Simpkins, S. H. Scholle, and K. J. Kelleher. 2004. "Parent/caregiver satisfaction with their child's Medicaid plan and behavioral health providers." *Mental Health Serv Res* 6(1):23-32.

Kessler, R. C., C. Barber, H. G. Birnbaum, R. G. Frank, P. E. Greenberg, R. M. Rose, G. E. Simon, and P. Wang. 1999. "Depression in the workplace: effects on short-term disability." *Health Aff (Millwood)* 18(5):163-71.

Kim, M., A. M. Zaslavsky, and P. D. Cleary. 2005. "Adjusting Pediatric CAHPS scores to ensure fair comparison of health plan performances." *Med Care* 43(1):44-52.

McGlynn, Elizabeth A. et al. 2003. "The Quality of Health Care Delivered to Adults in the

United States." *The New England Journal of Medicine* 348(26): 2635–2645.

NCBD. 2005. "National CAHPS Benchmarking Database" [accessed on September 1, 2005]. Available at: www.cahps.ahrq.gov.

NCQA. 2000. "Health Plan Employer Data and Information Set 2000, Volume 2: Technical Specifications". Washington, D.C.: National Committee for Quality Assurance.

NCQA. 2005. "Measuring the Quality of America's Health Care" [accessed on 11/30/05, 2005]. Available at: http://www.ncqa.org.

O'Malley, A. J., A. M. Zaslavsky, M. N. Elliot, L. Zaborski, and P. D. Cleary. 2005. "Case-Mix adjustment of the CAHPS® Hospital survey responses." *Health Serv Res* 40(6 (Pt 2)):2078-95.

Pincus, H. A. and A. R. Pettit. 2001. "The societal costs of chronic major depression." *J Clin Psychiatry* 62 Suppl 6:5-9.

Rohland, B. M., D. R. Langbehn, and J. E. Rohrer. 2000. "Relationship between service effectiveness and satisfaction among persons receiving Medicaid mental health services." *Psychiatr Serv* 51(2):248-50.

Shaul, J. A., S. V. Eisen, B. Clarridge, V. Stringfellow, F. J. Fowler Jr, and P. Cleary. 2001. "Experience of Care and Health Outcomes (ECHO)™ Survey Field Test Report: Survey Evaluation.". Boston, MA: Harvard Medical School, Dept of Health Care Policy.

Wang, P. S., G. Simon, and R. C. Kessler. 2003. "The economic burden of depression and the cost-effectiveness of treatment." *Int J Methods Psychiatr Res* 12(1):22-33.

Wells K., C. Sherbourne, M. Schoenbaum, S. Ettner, N. Duan, J. Miranda, J. Unutzer, L.V. Rubenstein. 2004. "Five-Year Impact of Quality Improvement for Depression: Results of a Group-Level Randomized Controlled Trial." *Archives of General Psychiatry*, 61:378-386.4

Zaslavsky, A. M. 1998. "Issues in case-mix adjustment of measures of the quality of health plans". Proceedings, Government and Social Statistics Sections, American Statistical Association.

Zaslavsky, A. M. 2001. "Statistical issues in reporting quality data: small samples and casemix variation." *Int J Qual Health Care* 13(6):481-88.

Zaslavsky, A. M., L. B. Zaborski, L. Ding, J. A. Shaul, M. J. Cioffi, and P. D. Cleary. 2001. "Adjusting performance measures to ensure equitable plan comparisons." *Health Care Finan Review* 22(3):109-26.

2b4.2. 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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

See 2b4.1

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

Of the 14,482 individuals in the initial sample, 18 percent were ineligible to receive the survey because of incorrect contact information, language barriers, illness, or death. The percentage of ineligible recipients ranged from 2 to 62 percent across the six sites and was generally lower among commercial plans (2 to 34%) than Medicaid plans (25 to 62%). Among the 11,855 eligible survey recipients, 48 percent responded. Among the 5,671 respondents, 28 percent were excluded from the analysis because they reported they were no longer enrolled in the plan, had been enrolled for less than the survey period of one year, had not received behavioral health services in the past 12 months as administrative records had indicated, or had someone else complete the survey for them (respondents who reported only receiving some help with reading, writing, or translation were not excluded). The remaining 3,067 enrollees in commercial plans and 1,001 enrollees in Medicaid plans (4,068 total) who reported receiving behavioral health services within the past 12 months comprised the analysis sample.

Data were available to compare respondents and non-respondents in three of the six sites. Respondents were significantly more likely to be older and to be female than non-respondents (data not shown). In Minnesota respondents had significantly more mental health visits than non-respondents. In New Jersey respondents were significantly more likely to have an alcohol/drug-related visit. No other significant differences between respondents and non- respondents were found.

In models that only controlled for plan effects, self-reported general health status and mental health status had positive, statistically-significant associations with global ratings and with nearly all of the report questions in the commercial and Medicaid samples. When tested simultaneously, mental health status was statistically significant more often than general health status, but general health status remained significant in a substantial number of the models and sometimes was significant when mental health status was not (results not shown). Consequently, both health status measures were retained for the base model.

After controlling for mental health, general health, and plan effects, older age was frequently associated with more positive ratings and reports, while having more education was often associated with less positive ratings and reports. Although not significantly related to either of the global ratings, alcohol/drug use was frequently associated with reports about experiences, sometimes positively and sometimes negatively. Significant associations with gender and race/ethnicity were generally less frequent, although Black race/ethnicity was negatively associated with several reports in the commercial sample. Income was a significant predictor of plan ratings and several reports about experiences in the commercial sample.

We next examined whether the linear specification of case mix adjuster variables was appropriate. We find that the increase in adjusted R-squared from using a categorical specification is modest. Considering the global rating of behavioral health care, the linear specification performs the worst for education, with a 15 percent increase in explanatory power (EP) from a categorical specification. Considering the global rating of behavioral health plan, in three of five ordinal variables, the linear specification has a higher adjusted R-squared than the categorical specification. The linear specification of income performs somewhat poorly relative to the categorical specification, with an increase in adjusted R-squared from .0236 to .0363. Given the generally modest improvement in model fit due to categorical specifications, the desire to keep the model parsimonious to allow for plan interactions, and the desire to have a model as consistent with other CAHPS adjustment models when appropriate, we choose to keep the ordinal specification in the model.

In models that controlled only for plan effects, mental health status had the highest levels of EP for both the ratings and reports in the commercial and Medicaid samples with one exception. The mean EP levels for Hispanic ethnicity were higher in both samples, but Hispanic ethnicity was related to considerably fewer report questions compared to mental health status so was not chosen for the base model. General health status, the other variable significantly related to nearly all the ratings and reports, had high mean levels of EP compared to most other variables, so it was also chosen along with mental health status for the base model.

After controlling for variables in the base model, education had sizeable EP for ratings in both samples. Additionally, age and Black race/ethnicity were important in the Medicaid sample, and income was important in the commercial sample. With respect to reports, Hispanic ethnicity had the highest mean EP in both the commercial and Medicaid samples, followed by age in the Medicaid sample and income in the commercial sample. Although the mean EP value for age was low in the commercial sample, age was significantly related to half of the report questions. Alcohol/drug use had the third highest mean EP level in both the Medicaid and commercial samples. Gender was important in the Medicaid sample, but only modestly influential in the commercial sample. Education had relatively low levels of mean EP, but was significantly associated with a relatively high number of reports in each sample. For reports, Black race/ethnicity was more important than age and gender in the commercial sample, and more important than education in the Medicaid sample.

The final adjustment model included mental and general health status, education, age, Black race, Hispanic ethnicity, and alcohol/drug use. Alternative models adding income for the commercial sample and gender for the Medicaid sample were also evaluated.

On average, the effects of the case mix adjustment on plan scores generally appear to be modest. For example, for the global rating of behavioral health care, which has the largest adjustment, the mean absolute adjustment was .08 in the commercial sample and .10 in the Medicaid sample. For some plans, adjustment is important. For example, the largest positive adjustment was 0.2, which is equivalent to the mean difference between commercial and Medicaid plans. When considering changes in plan rankings, the effect was greater, although still modest. Most Kendall tau correlations were above 0.80, indicating that adjustment altered plan rankings in less than 10 percent of the possible pairings between plans ((1-0.80)/2). Slightly greater percentages of plan pairs (13 to 14 percent) switched their internal ranked order for *the Global Rating of Behavioral Health Plan* and *Plan Approval and Service* in the Medicaid sample.

Adding gender to the model for the Medicaid plans resulted in no substantial change in the average magnitude of adjustments for the ratings and report composites (results not shown). However, changes in the largest positive and negative adjustments led to an increase in the percentage of pairs that switched order for the *Global Rating of Behavioral Health Care* (from 0 to 5 percent) and the *Global Rating of Behavioral Health Plan* (from 13 to 20 percent). Similarly, adding income to the model did not have much impact on scores for the commercial plans. The mean absolute adjustment and the largest positive and negative adjustments were generally greater with income in the model. However, the adjustments increased the percentage of plan pairs that switched order only in the case of *Timely Access to Care*, and the increase was modest (from 0 to 5 percent).

Plan interactions with mental and general health status were statistically significant in the models predicting *Global Rating of Behavioral Health Car*e, and only with mental health status for *Global Rating of Behavioral Health Plan* (results not shown). In addition, seven plan-by-mental health interactions, and eight plan-by-general health interactions, were statistically significant in the models predicting the 22 report questions. In contrast, fewer plan interactions with the demographic characteristics were statistically significant—three with age, two with gender, two with Black race/ethnicity, four with Hispanic race/ethnicity, two with education, and two with income. In 24 tests, one or two interactions would be expected by chance.

If we assume that the effect of health status on individual reports is consistent across plans independent of actual experience with the plan, the significant interaction effects of plan and health status imply that there are quality differences by health plans. To illustrate the nature of plan differences, we stratified plan scores for *Global Rating of Behavioral Health Care* by mental health status (Figure 2). In nearly all plans, individuals self-reporting excellent or very good mental health status rated their care more highly than individuals self-reporting fair or poor mental health status. However, the difference between the average rating given by respondents in the "poor/fair" category and respondents in the "very good/excellent" category ranged from 0.0 to 2.6 points on the 0 to 10 response scale. To put this in context, the overall average rating was about 7.9. Mean ratings among the sickest respondents varied more across plans than mean ratings among the healthiest respondents (standard deviation = 0.72 and 0.29, respectively).

Summary (references in section 2b4.2)

Consistent with other CAHPS® studies (Kim et al. 2005; O'Malley et al. 2005; Zaslavsky et al. 2001), the average impact of case mix adjustment on plan scores for ratings and reports collected from behavioral health care patients was modest. Adjustments did change plan rankings in a few cases for both the commercial and Medicaid plans, with adjustments typically being larger for the Medicaid plans. For a few individual plans, the change in some summary scores was large.

Although the effects in these data are modest, adjustments would be larger in groups of plans with greater inter-plan heterogeneity in patient characteristics. Whether the impact is large or small, case mix adjustment may still be important to maintain the credibility of patient reports as a quality metric. In the absence of case mix adjustment, plans that believe their patients have worse health status than patients in other plans may believe summary scores are suspect and be reluctant to rely on them as a quality indicator. The fact that case mix adjustment can have a meaningful effect on plan scores and rankings, and requires only a small amount of information that is typically collected for other purposes such as subgroup analyses, makes it worthwhile to carry out. Doing so can preserve the face validity of the results for plans who might otherwise complain that their patients are more severe than is typical.

As expected, the self-reported health status measures were the strongest and most consistent predictors of ratings and reports among the personal characteristics included in this study. Mental health status was frequently a strong predictor, as in other studies (Zaslavsky et al. 2001), although general health status remained important in several cases after controlling for mental health status. This association may be due to general reporting tendencies that are associated, for instance, with general life satisfaction (Rohland et al. 2000) or with effects of mental illness on mood and perception. Patients in worse mental health may also receive lower quality of care than patients in better mental health. Providers are likely to have more difficulty communicating with patients who are distressed and may tend to unconsciously convey negative attitudes and behaviors towards patients in poor mental health (Hall et al. 1993). The chronic and recurring nature of many behavioral health conditions, and the uncertainty involved in determining the best treatment strategy for a particular patient, increase the likelihood that multiple treatment approaches will be attempted before symptoms are relieved, which may cause dissatisfaction with care.

As in other studies, older respondents tended to give more positive ratings and reports than younger respondents (Zaslavsky et al. 2001). Older respondents may have lower expectations regarding their care and/or more respect for providers, leading them to give more positive reports and ratings. Older patients also may receive better care than younger patients if, for instance, providers tend to be more attentive to older patients (Cleary et al. 1992). Education was negatively associated with positive ratings. Because it is unlikely that respondents with more education receive worse care than respondents with less education, the findings are consistent with the hypothesis that highly educated respondents have higher expectations regarding their care which result in less favorable assessments. The mixed

findings related to income are consistent with other studies (Dow et al. 2001; Heflinger et al. 2004). Although a substantial number of respondents are reluctant to provide income data, an income item with coarse categories may be useful for case mix adjustment when the income distributions of plan populations vary greatly. Because the effects of race/ethnicity were not the same in commercial and Medicaid plans, it may be important to estimate separate models for the two types of plans (as we do here) and make comparisons only among plans of the same type.

The regression models used to control for case mix differences assume that the effects of adjusters are equal within each plan. While the effects of most demographic adjusters were similar across plans, the effects of mental health status varied across plans for several rating and report questions. Therefore, the adjusted plan scores, and possibly the rankings of plans based on those scores, would depend on whether the plans were compared with respect to how patients with poor, average, or excellent mental health status would rate them (Zaslavsky 1998). Reporting separate summaries for patients in relatively poor and good health (provided that sample sizes for these groups were sufficient within each plan) may be important for identifying performance differences among plans.

In summary, mental health status, general health status, alcohol/drug use, age, education, and race/ethnicity were identified as relevant case mix adjusters for the ECHO™ survey, although the case mix adjustment model resulted in only minor changes to plan ratings and rankings.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

N/A

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

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

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

Most of the ECHO Survey items are able to discriminate between behavioral health plans. Of the 47 items evaluated with analysis of variance, all but 8 detected statistically significant differences between organizations. The statistically significant items had F-statistics ranging from 2.25 to 30.16 (Shaul, Eisen et al. 2001).

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

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

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

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

The main way of minimizing the effects of unit (as opposed to item) missing data is to maximize response rates. The AHRQ web site includes guidance on how to do that. In addition, since the patient characteristics used in the case mix adjustment model are related to the probability of response, case-mix adjusting the unit means reduces much of the potential inter-unit bias due to non-response.

For item non-response, the CAHPS Consortium has developed detailed instructions regarding how to deal with missing data:

http://www.ahrq.gov/professionals/quality-patient-safety/patientsafetyculture/pharmacy/toolkit/pharmsopsuserguide5.html

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
See 2b5

Below is some information about missing data in two published studies describing the development and psychometric characteristics of the ECHO instrument.

Shaul JA, Eisen SV, Stringfellow VL, Clarridge BR, Hermann RC, Nelson D, Anderson E, Kubrin AI, Leff HS, Cleary PD. Use of consumer ratings for quality improvement in behavioral health insurance plans. Jt Comm J of Qual Imp; 2001; 27: 216-229.

In 1998 and 1999, data were collected from five groups of adult patients in commercial health plans and five groups of adult patients in public assistance health plans with services received through four organizations, one of three MBHOs or a health system. Patients who received behavioral health care services during the previous year were mailed the CABHS survey. Non-respondents were contacted by telephone to complete the survey.

Response rates ranged from 49% to 65% for commercial patient groups and from 36% to 51% for public assistance patients.

Eisen, S.V., B. Clarridge, V. Stringfellow, J.A. Shaul, and P.D. Cleary. "Toward a National Report Card: Measuring Consumer Experiences with Behavioral Health Services. In B. Dickey and L. Sederer (Eds.), Improving Mental Health Care: Commitment to Quality. Washington, DC: APA Press, 2001. Chapter 9;115-134.

The CABHS survey was field tested with two groups of mental health consumers: commercially insured individuals (N=200), and Medicaid enrollees (N=300). Both groups were members of an HMO for which the behavioral health component was managed by an external managed behavioral health care organization. The survey was conducted by a survey research center that was independent of the HMO and the managed care organization.

Survey response rates reached an acceptable level (63%) for commercially insured consumers who were reachable by telephone. Response rates to mailed surveys alone were about 32%, highlighting the importance of telephone follow-up. A large percentage of consumers, particularly Medicaid enrollees, were not reachable by phone (55%). Of those reached by telephone, 57% participated in the survey.

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

The CAHPS consortium has done extensive research on how best to deal with missing item data and how to maximize response rates. There also has been extensive research on how best to adjust results for survey respondent characteristics. Although those case-mix strategies are primarily designed to account for differences in response tendencies, several of the characteristics in the models are related to the probability of responding (e.g. age and sex) and so the adjusted scores are less biased by missing data. In addition, research has shown that the characteristics of non-respondents vary by survey mode. That is, the individuals least likely to respond to mailed surveys are different than the characteristics of individuals least likely to respond to telephone surveys. The CAHPS Consortium recommends mixed mode (mail and phone) surveys in part because of this and to increase response rates.

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.**
Other
If other: Survey measures

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**.
No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
ECHO is based on patient reported survey data. The CAHPS consortium is doing research on using patient portals to collect such data and currently is conducing a study in a large practice in Massachusetts of such an approach. If that study is successful, we probably will do further research on collecting patient reported data through patient portals. One barrier to doing this, however, is that many organizations providing behavioral health services do not have patient portals and/or the data collected electronically is subject to protections that make it difficult to use such information for assessing the quality of patient centered care. The consortium continues to study this issue however.

Another approach to collecting patient-reported data that the CAHPS Consortium is actively studying is the use of electronic devices (e.g., PDAs/Smartphones, tablets). Unfortunately,these approaches have not been refined to the point that the data collected is adequately representative of the patient population to which inferences are to be made.

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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
ECHO data collection follows established CAHPS survey protocols:

http://www.ahrq.gov/cahps/surveys-guidance/echo/index.html

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)*.
ECHO surveys and survey materials, including the MACRO for analyses are publicly available at no cost on the AHRQ Web Site.

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)
Use Unknown	

4a.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

The CAHPS Consortium makes surveys, including the ECHO surveys, available to the public for use but does not systematically track use.

We have three indications, however, of on-going use, downloads from the CAHPS-ECHO Web Site, ECHO survey web page visits, and requests for technical assistance. Westat serves as a public resource for CAHPS users and one of its functions is to answer requests about specific instruments. To provide support for our application for maintenance of the ECHO endorsement, we compiled selective data on these two indicators.

Westat examined external requests for technical assistance specifically related to the ECHO surveys over the past two years. In 2015 they received 83 inquiries and as of December 15, 2016 they had received 33 inquiries in 2016, for a total of 116 in the past two years.

In 2015, there were approximately 250 ECHO web page visits a month. In 2016, the number of ECHO web page visits rose fairly steadily form about 250 a month to 650 a month by the end of the year.

In the last six months of 2015 there were approximately 40 downloads a month of ECHO surveys and in 2016 the number of instrument downloads a month varied between about 60 and 120 a month.

These numbers of downloads and inquiries likely represent only a small subset of the number of organizations and entities (e.g. state Medicaid programs) using ECHO, but at least they indicate on-going interest in use.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The CAHPS Consortium makes surveys, including the ECHO surveys, available to the public for use but does not systematically track use.

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

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

4b. **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 CAHPS Consortium makes surveys, including the ECHO surveys, available to the public for use but does not systematically track use. We have anecdotal information that states and plans are using ECHO data to stimulate and monitor improvement, but we do not systematically compile such data.

4c. Unintended Consequences

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

4c.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
None

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
None

4d1.1. **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.
See response to 4b.

4d1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
See response to 4b.

4d2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
See response to 4b.

4d2.2. **Summarize the feedback obtained from those being measured.**
NA

4d2.3. **Summarize the feedback obtained from other users**
NA

4d.3. **Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
NA

5. Comparison to Related or Competing Measures

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

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

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

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

5a. Harmonization 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?

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.)

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.
Available at measure-specific web page URL identified in S.1 **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Agency for Healthcare Research and Quality
Co.2 **Point of Contact:** Caren, Ginsberg, caren.ginsberg@AHRQ.HHS.Gov, 301-427-1412-
Co.3 **Measure Developer if different from Measure Steward:**
Co.4 **Point of Contact:**

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 ECHO Development Group comprised representatives from NCQA, NCQA's Behavioral Health Measurement Advisory Panel, the CAHPS consortium, the MHSIP development group, the Human Services Research Institute, the Center for Mental Health Services, the Forum on Performance Measures for Behavioral Health and Related Service Systems, the Washington Circle Group, the American Managed Behavioral Healthcare Association (AMBHA), and the National Alliance for the Mentally Ill participated in the development of the ECHO survey. Consumers, clinicians and behavioral health policy experts were also members of the survey development team. Consumers played several important roles in the development of the ECHO survey. They provided feedback about the content and design of the instrument, participated in focus groups and one-on-one interviews, and were part of the survey's development team. The information provided by consumers about the content and design of the ECHO survey has been crucial to the survey's ability to ask about concepts that are important to consumers for the evaluation of their behavioral health treatment and for which they are the best or only source of information.

Individuals and members of groups that contributed to the development of the ECHO measures.

Thomas M. Achenbach, Ph.D.
University of Vermont
Burlington, VT

Brian V. Abbott
Texas A&M University
College Station, TX

Ross B. Andelman, M.D.
Contra Costa Children's Mental Health Services
Concord, CA

Robert P. Archer, Ph.D.
Eastern Virginia Medical School
Norfolk, VA

C. Clifford Attkisson, Ph.D.
University of California, San Francisco
San Francisco, CA

Steven E. Bailley
University of Texas - Houston Health Sciences Center
Houston, TX

Thomas Beers, Ph.D.
Kaiser Permanente San Diego Chemical Dependency Program
San Diego, CA

Albert J. Bellanger
Harvard Medical School
Cambridge, MA

Larry E. Beutler, Ph.D.
Pacific Graduate School of Psychology
Palo Alto, CA

Phillip J Brantley, Ph.D.
Pennington Biomedical Research Center
Baton Rouge, LA

Gary M. Burlingame
Brigham Young University
Provo, UT

James N. Butcher, Ph.D.
University of Minnesota
Minneapolis, MN

David L. Carlston,Ph.D.
Ohio University
Athens, OH

Antonio Cepeda-Benito, M.D.
Texas A&M University
College Station, TX

Dianne L. Chambless, Ph.D.
University of Pennsylvania
Philadelphia, PA

James A. Ciarlo, Ph.D.
University of Denver
Denver, CO

Paul D. Cleary, Ph.D.
Harvard Medical School
Boston, MA

James R. Clopton, Ph.D.

Texas Tech University
Lubbock, TX

John D. Cone, Ph.D.
Alliant International University
San Diego, CA

C. Keith Conners, Ph.D.
Duke University
Durham, NC

Jonathan C. Cox
Brigham Young University
Provo, UT

William J. Culpepper, M.S.
University of Maryland
Baltimore, Maryland

Constance J. Dahlberg
Alliant International University
San Diego, CA

Allen S. Daniels, Ed.D.
Alliance Behavioral Care
University of Cincinnati
The American Managed Behavioral Healthcare Association
Cincinnati, OH

Edwin de Beurs
Leiden University

Leonard R. Derogatis, Ph.D.
University of Maryland
Baltimore, MD

Kathy Dowell
Ohio University
Athens, OH

Gareth R. Dutton, M.A.
Louisiana State University
Baton Rouge, LA

William W. Eaton, Ph.D.
Johns Hopkins University
Baltimore, MD

Susan V. Eisen, Ph.D.
Edith Nourse Rogers Memorial Veterans Hospital
Boston University
Boston, MA

Jeffery N. Epstein, Ph.D.
Duke University
Durham, NC

Alex Espadas
University of Texas - Houston Health Sciences Center
Houston, TX

Laura E. Evison, CNM, MSN
University of Maryland
Baltimore, MD

Kya Fawley
Northwestern University
Chicago, IL

Maureen Fitzpatrick, CRNP, MSN
University of Maryland
Baltimore, MD

Jenny Fleming
University of California, Santa Barbara
Santa Barbara, CA

Michael B. Frisch, Ph.D.
Baylor University
Waco, TX

Anthony B. Gerard, PhD.
Western Psychological Services
Los Angeles, CA

Sona Gevorkian
Massachusetts General Hospital
Boston, MA

David H. Gleaves
Texas A&M University
College Station, TX

Pamela Greenberg, MPP
The American Managed Behavioral Healthcare Association

Roger L. Greene, Ph.D.
Pacific Graduate School of Psychology
Palo Alto, CA

Thomas K. Greenfield, Ph.D.
University of California, San Francisco
San Francisco, CA

Ann T. Gregersen

Brigham Young University
Provo, UT

Grant R. Grissom, Ph.D.
Polaris Health Directions, Inc.
Langhorne, PA

Seth D. Grossman, M.A., M.S.
IASPP
Carlos Albizu University
Miami, FL

Kurt Hahlweg
Technische Universitaet Braunschweig
Braunschweig, Germany

Steven R. Hahn, M.D.
Albert Einstein College of Medicine
Jacobi Medical Center
Bronx, NY

Ashley Hanson
University of Alabama
Tuscaloosa, AL

Nancy M. Hatcher
University of Georgia
Athens, GA

Derek Hatfield
Ohio University
Athens, OH

Eric J. Hawkins
Brigham Young University
Provo, UT

Jena Helgerson
Northwestern University
Chicago, IL

Kay Hodges, Ph.D.
Eastern Michigan University
Ann Arbor, MN

Elizabeth A. Irvin
Services Research Group, Inc.
Simmons College

Gary Jeager, M.D.
Kaiser Permanente Harbor City Chemical Dependency Program
Harbor City, CA

R. W. Kamphaus, Ph.D.
University of Georgia
Athens, GA

Jennifer Karpe
University of Alabama
Tuscaloosa, AL

Sangwon Kim
University of Georgia
Athens, GA

Scott H. Kollins, Ph.D.
Duke University
Durham, NC

Kenneth A. Kobak
Dean Foundation for Health Research and Education
Research Training Associates
Madison, WI

Teresa L. Kramer, Ph.D.
Centers for Mental Healthcare Research
University of Arkansas for Medical Sciences
Little Rock, AR

Kurt Kroenke, M.D.
Regenstrief Institute for Health Care
Indiana University School of Medicine
Indianapolis, IN

Samuel E. Krug, Ph.D.
MetriTech, Inc.
Champaign, IL

David Lachar, Ph.D.
University of Texas - Houston Health Sciences Center
Houston, TX

Michael J. Lambert, Ph.D.
Brigham Young University
Provo, UT

Jeanne M. Langraf, M.A.
HealthAct
Boston, MA

William W. Latimer
Johns Hopkins University
Baltimore, MD

Jean-Philippe Laurenceau
Texas A&M University

College Station, TX

John S. Lyons, Ph.D.
Northwestern University
Chicago, IL

Mary Malik
University of California, Santa Barbara
Santa Barbara, CA

John S. March, M.D.
Duke University Medical Center
Durham, NC

Mark E. Maruish, Ph.D.
Ingenix Pharmaceutical Services
Eden Prairie, MN

Sarah E. Meagher, M.S.
University of Miami
Miami, FL

Gregorio Melendez, Ph.D.
Ohio University
Athens, OH

Theodore Millon, Ph.D., D.Sc.
IASPP
Miami, FL

Carla Moleiro, Ph.D.
University of California, Santa Barbara
Santa Barbara, CA

Leslie C. Morey, Ph.D.
Texas A&M University
College Station, TX

Carles Muntaner

Jack A. Naglieri, Ph.D.
George Mason University
Fairfax, VA

Charles Negy
University of Central Florida

Frederick L. Newman, Ph.D.
Florida International University
Miami, FL

Sharon-Lise T. Normand
Harvard Medical School

Cambridge, MA

Benjamin M. Ogles
Ohio University
Athens, OH

Ashley E. Owen
University of South Florida
Tampa, FL

James D. A. Parker, Ph.D.
Trent University
Ontario, Canada

Julia N. Perry

Steven I. Pfeiffer, Ph.D.

James O. Prochaska, Ph.D.
University of Rhode Island
Kingston, RI

Janice M. Prochaska,. Ph.D.
Pro-Change Behavior Systems, Inc.

Leslie A. Rescorla,
Bryn Mawr College

Eric C. Reheiser
University of South Florida
Tampa, FL

Cecil R. Reynolds, Ph.D.
Texas A&M University
College Station, TX

William M. Reynolds
Humboldt State University

James M. Robbins, Ph.D.
Centers for Mental Healthcare Research
University of Arkansas for Medical Sciences
Little Rock, AR

Abram B. Rosenblatt, Ph.D.
University of California, San Francisco
San Francisco, CA

Douglas Rugh
Florida International University
Miami, FL

Scott Sangsland, M.A.

Kaiser Permanente,
Southern California Permanente Medical Group

Forrest Scoggin, Ph.D.
University of Alabama
Tuscaloosa, AL

James A. Shaul, MHA
Harvard Medical School
Boston, MA

Gill Sitarenios, Ph.D.
Multi-Health Systems, Inc
Toronto, ON, Canada

Corey Smith

G. Richard Smith, M.D.
University of Arkansas for Medical Sciences
Little Rock, AR

Douglas K. Snyder, Ph.D.
Texas A&M University
College Station, TX

Charles D. Spielberger, Ph.D.
University of South Florida
Tampa, FL

Robert L. Spitzer, M.D.
New York State Psychiatric Institute
Columbia University
New York, NY

Steven Stein, Ph.D.
Multi-Health Systems, Inc
Toronto, ON, Canada

Randy Stinchfield, Ph.D.
University of Minnesota
Minneapolis, MN

Sumner J. Sydeman
Northern Arizona University
Flagstaff, AZ

Elana Sydney, M.D.
Albert Einstein College of Medicine
Jacobi Medical Center
Bronx, NY

Hani Talebi
University of California, Santa Barbara

Santa Barbara, CA

Manuel J. Tejeda
Barry University

Allen Tien

John E. Ware, Jr., Ph.D.
Tufts Medical School
QualityMetric, Inc.
Lincoln, RI

Irving B. Weiner, Ph.D.
University of South Florida
Tampa, FL

M. Gawain Wells
Brigham Young University
Provo, UT

Douglas Welsh
University of Alabama
Tuscaloosa, AL

Janet B. W. Williams, D.S.W.
New York State Psychiatric Institute
Columbia University
New York, NY

Kimberly A. Wilson
Stanford University
Palo Alto, CA

Ken C. Winters. Ph.D.
University of Minnesota
Minneapolis, MN

Stephen E. Wong, Ph.D.
Florida International University
Miami, FL

Karen B. Wood, M.A.
Louisiana State University
Baton Rouge, LA

Michele Ybarra

Daniels AS, Shaul JA, Greenberg P, Cleary PD. "The Experience of Care and Health Outcomes Survey (ECHO): A Consumer Survey to Collect Ratings of Behavioral Health Care Treatment, Outcomes and Plans." In: M.E. Maruish (Ed), The Use of Psychological Testing for Treatment Planning and Outcomes Assessment. Third Edition. Vol. 3 Instruments for Adults. Fairfax, VA: Lawrence Erlbaum Assoc., 2004. Chapter 29, 839-866.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2007
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?** As needed as determined by major CAHPS revisions and changes in behavioral health standards
Ad.5 **When is the next scheduled review/update for this measure?** 03, 2017

Ad.6 **Copyright statement:** AHRQ holds the Trademark to ECHO
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:** The organization affiliation of Caren Ginsberg should be the Agency for Healthcare Research and Quality, not CMS. She should be listed as the Steward (I coudl not change these settings on-line; Paul).

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0027
Corresponding Measures:
Measure Title: Medical Assistance With Smoking and Tobacco Use Cessation
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The three components of this measure assess different facets of providing medical assistance with smoking and tobacco use cessation:

Advising Smokers and Tobacco Users to Quit: A rolling average represents the percentage of patients 18 years of age and older who are current smokers or tobacco users and who received advice to quit during the measurement year.

Discussing Cessation Medications: A rolling average represents the percentage of patients 18 years of age and older who are current smokers or tobacco users and who discussed or were recommended cessation medications during the measurement year.

Discussing Cessation Strategies: A rolling average represents the percentage of patients 18 years of age and older who are current smokers or tobacco users and who discussed or were provided cessation methods or strategies during the measurement year.
Developer Rationale: Tobacco smoking is the leading cause of preventable disease and death in the United States, resulting in approximately 480,000 premature deaths and more than $300 billion in direct health care expenditures and productivity losses each year (HHS, 2014). Premature deaths due to smoking, including deaths from lung cancer, pulmonary diseases, coronary heart disease, pregnancy concerns, and residential fires, numbered over 20 million between 1965 and 2014 (HHS, 2014). Although the consumption of cigarettes continues to decline (with a decrease from 20.9 percent in 2005 to 16.8 percent in 2014 (Centers for Disease Control and Prevention, 2015)), the use of electronic cigarettes, or e-cigarettes, has more than doubled between 2011 and 2012, especially among adolescents (HHS, 2014).

The strongest evidence on increasing smoking cessation comes from studies involving physician or nurse's advice, tailored self-help materials, or telephone counseling (Siu, 2015). For example, interventions that involve physician or nurse advice are associated with smoking abstinence at six months or more after the intervention (8.0 percent for physicians and 13.3 percent for nurses) compared to no advice or usual care (4.8 percent for physicians and 11.3 percent for nurses). Study participants who receive tailored self-help materials are more likely to cease smoking at six months or more when compared to study participants who did not receive self-help materials (7.1 percent vs. 5.8 percent). The U.S. Prevention Services Task Force (USPSTF) has found evidence that smoking cessation decreases the risk for heart disease, lung disease, and stroke through a review of published literature. The USPSTF also highlights evidence that smoking and tobacco use cessation interventions (including counseling sessions and pharmacotherapy)

are effective in increasing the proportion of patients who remain tobacco-free for at least 6 months to 1 year depending on length of intervention (Siu, 2015).

Citations:
Centers for Disease Control and Prevention. (2015). Current Cigarette Smoking Among Adults—United States, 2005–2014. Morbidity and Mortality Weekly Report;64(44):1233–40. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6444a2.htm

Siu, A. L. (2015). Behavioral and Pharmacotherapy Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Women: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine, 163(8), 622-635. Retrieved from: http://annals.org/aim/article/2443060/behavioral-pharmacotherapy-interventions-tobacco-smoking-cessation-adults-including-pregnant-women

U.S. Department of Health and Human Services (HHS). (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Retrieved from https://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf.

Numerator Statement: Advising Smokers and Tobacco Users to Quit:
Patients who indicated that they received advice to quit smoking or using tobacco from their doctor or health provider

Discussing Cessation Medications:
Patients who indicated that their doctor or health provider recommended or discussed smoking or tobacco cessation medications

Discussing Cessation Strategies:
Patients who indicated their doctor or health provider discussed or provided smoking or tobacco cessation methods and strategies other than medication
Denominator Statement: Patients 18 years and older who responded to the CAHPS survey and indicated that they were current smokers or tobacco users during the measurement year or in the last 6 months for Medicaid and Medicare.
Denominator Exclusions: None

Measure Type: Process
Data Source: Patient Reported Data
Level of Analysis: Health Plan, Integrated Delivery System

Original Endorsement Date: Aug 10, 2009 **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 *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.

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 201, the developer provided evidence from the USPSTF (tobacco-related counseling in adults and pregnant women), the Institute for Clinical Systems Improvement (tobacco use prevention and cessation for adults and mature adolescents), the VA/DoD (clinical practice guideline for the management of tobacco use), and the Public Health Service (treating tobacco use and dependence).The committee agreed that this measure met the importance criteria, but expressed concern for recall bias due to the length of time between patient-physician interaction and when the survey was distributed.

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 a 2015 guideline from the USPSTF on behavioral and pharmacotherapy interventions for tobacco smoking cessation in adults, including pregnant women.
    - The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and U.S. Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. **(Grade A recommendation)**
    - The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. **(Grade A recommendation)**
    - The USPSTF concludes that the current evidence is **insufficient** to assess the balance of benefits and harms of pharmacotherapy interventions for tobacco cessation in pregnant women. (I statement)
    - The USPSTF concludes that the current evidence is **insufficient** to recommend electronic nicotine delivery systems (ENDS) for tobacco cessation in adults, including pregnant women. The USPSTF recommends that clinicians direct patients who smoke tobacco to other cessation interventions with established effectiveness and safety (previously stated). (I statement)

Exception to evidence

NA

Questions for the Committee:

- *The evidence provided by the developer is updated and directionally the same as 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

Patient-reported process measure based on systematic review (Box 3)→QQC presented (Box 4)→Quantity: high; Quality: high; Consistency: high (Box 5)→High (Box 5a)→High

The highest possible rating is 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.
NOTE: Data for this measure is obtained from the CAHPS Health Plan Surveys.
Performance data are summarized at the health plan level (commercial, Medicare, Medicaid) for 2014-2016 for each of the 3 rates reported in this measure. Data are summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare beneficiaries.

Plan	Year	Mean	Standard Deviation	10th Quartile	90th Quartile
Advising smokers to quit					
Medicare	**2014**	84%	7%	75%	92%
	2015	86%	6%	78%	93%
	2016	86%	6%	78%	93%
Commercial (rolling average)	**2014**	74%	8%	66%	85%
	2015	75%	7%	64%	83%
	2016	75%	7%	66%	83%
Medicaid (rolling average)	**2014**	76%	5%	69%	81%
	2015	76%	6%	68%	82%
	2016	76%	6%	68%	82%
Discussing cessation medications					
Commercial	**2014**	49%	9%	37%	61%
	2015	49%	8%	38%	58%
	2016	48%	8%	41%	61%
Medicaid	**2014**	47%	8%	38%	57%
	2015	47%	8%	36%	57%
	2016	48%	8%	37%	58%
Discussing cessation strategies					
Commercial	**2014**	43%	10%	31%	57%
	2015	44%	9%	32%	56%
	2016	44%	9%	34%	58%
Medicaid	**2014**	42%	7%	34%	51%
	2015	42%	7%	34%	51%
	2016	43%	7%	34%	52%

Disparities

HEDIS measures are stratified by type of insurance, but not stratified by race. They note significant disparities in tobacco use among certain populations, but provided limited evidence on the disparities among smoking cessation efforts in these populations.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is a discussion needed regarding the relatively minimal change in these rates over the last 3 years?*
- *Are you aware of evidence that disparities exist in smoking cessation efforts?*

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)

1.a. Evidence to Support Measure Focus

Comments:

**Updated evidence is provided - 2015 USPSTF guideline on behavioral ad pharmacotherapy interventions for tobacco cessation. For adults the USPSTF recommendation is Grade A but insufficient evidence is claimed for recommendation related to pregnant women. This reviewer rates the evidence as HIGH and does not think there is any reason to further discuss and vote on evidence. Evidence rating for this reviewer is HIGH.

**Clearly defined and updated. Clear causal pathway.

** As discussed in the PA, the evidence provided by the developer is updated and directionally the same as that for the previous NQF
review, there is no need for repeat discussion and vote on Evidence.

** Health care provider screens adults for tobacco use >>> Identifies adults who are using tobacco >>> Advises adults who use tobacco to stop using tobacco and discusses cessation medications and strategies >>> Reduction in tobacco use >>> Improved health >>> Improved health outcomes (including mortality). Cites literature supporting efficacy of counseling and medication interventions.

** The evidence base supporting the fact that health care interventions is strong (and has probably only strengthened since this work was last updated.) Similarly, there is no doubt that smoking cessation improves health outcomes. I see no reason to revote the importance criterion.

1.b. Performance Gap

Comments:

** For this reviewer, the performance gap data for advising smokers to quit at the health plan level for commercial plans and Medicaid remains HIGH, less so for Medicare. for discussing cessation medications and strategies, performance gap is even higher and warrants a national performance measure. Might be useful for the developer to assess disparities among subpopulations by race and/or ethnicity.

**Despite ongoing work, there is still a gap particularly in parts of the recommended care as assessed.

** Performance data is provided and demonstrates a gap. There has been minimal change over the past 3 years. Of the included reviews that assessed specific subpopulations, the reviews were widely varied and included multiple review categories; one review concentrated on smokeless tobacco users, four reviews focused on cessation interventions for racial and ethnic minority groups, one only included results for young adults and one focused on interventions among older adult smokers. None of these reviews were considered part of the primary review. Of the included reviews that assessed pregnant women, there was considerably more evidence available on the effects of behavioral interventions during pregnancy than for pharmacotherapies.

Data is stratified by type of insurance (commercial, Medicaid, Medicare).

** There are continuing gaps in performance (that are a bit surprising given the importance of the topic and the relative simplicity of the intervention).

1.c. Composite
Comments:
**Yes

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Patient-reported data (CAHPS Health Plan Survey 5.0H, Adult Version; Medicare CAHPS)

Specifications:

- This measure is specified for the health plan level of analysis in the clinician office/clinic setting. The developer also notes that in addition to clinician visits, some respondents may recall contacts with an "other health professional" (wording used in survey question), which may include contacts with nurses or health plan staff.
- A higher score indicates better quality.
- This measure includes three rates:
  - Advising smokers and tobacco users to quit,
  - Discussing cessation medications, and
  - Discussing cessation strategies.
- In each rate, the numerator is identified by a specific question on the CAHPS survey. Numerator details (including specific questions) are provided for Medicare, Medicaid, and commercial plans.
- For all three rates, the denominator is patients 18 years and older who responded to the CAHPS survey and indicated that they were current smokers or tobacco users during the measurement year, or in the last 6 months for Medicaid and Medicare. Patients must answer both questions about current use as well as specific questions related to the numerator.
- Denominator details are provided for Medicare, Medicaid, and commercial plans.
- There are no exclusions.
- A calculation algorithm is provided.
- The developer states that a systematic sampling method is used for the CAHPS survey. For HEDIS/CAHPS surveys, basic instructions are given, including required sample size. For Medicare CAHPS, CMS provides the sample. NCQA evaluates a health plan's prior year's survey results to establish required sample sizes. The sample sizes are calculated to ensure a large enough sample to achieve 411 completed surveys per health plan and the developer provides suggested based on health plan.
- Proxy responses are not allowed for adult CAHPS survey.

Questions for the Committee:
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:
Previous measure score reliability testing included use of a signal to noise analysis for 2010 HEDIS data.

Describe any updates to testing:
The developer provided a 2016 update of measure score reliability testing for commercial and Medicaid plans.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Developer assessed measure score reliability using data from all health plans that submitted HEDIS data to NCQA for this measure and had a valid rate in 2015/2016. The developer provides data on the denominator sizes for reliability testing, including the number of plans and the median eligible number of patients per plan.
- NOTE: An underlying assumption regarding instrument-based performance measures is that they are based on instruments that have been shown to be both reliable and valid. Recently, NQF has operationalized this assumption by requiring demonstration of the reliability and validity of the underlying survey (i.e., data element reliability and validity). A short summary of the reliability of the CAHPS Health Plan surveys will be required in the future).

Results of reliability testing

Beta-binomial statistic for each measure rate

	Commercial		Medicaid		Medicare	
Year	**2010**	**2016**	**2010**	**2016**	**2010**	**2016**
Advising Smokers to Quit	0.62	0.69	0.61	0.75	0.95	Not updated
Discussing Cessation Medications	0.47	0.72	0.85	0.83	N/A	N/A
Discussing Cessation Strategies	0.70	0.77	0.79	0.77	N/A	N/A

The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Generally, a reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities.

Questions for the Committee:

- *Does the Committee think there is a need to re-vote on reliability?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Specifications are precise (Box 2)→empirical reliability testing (Box 4)→score level testing (Box 5)→signal-to-noise analysis shows moderate signal strength (Box 6)→Moderate

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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

For maintenance measures, summarize the validity testing from the prior review:
In 2011, the developer reported systematic assessment of face validity and basic information about cognitive testing of the CAHPS survey instrument done in 2008.

Describe any updates to validity testing:
In 2016, results of new construct validity testing are provided as well as additional information related to the 2008 cognitive testing of the instrument.

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☐ **Data element testing ~~against a gold standard~~** ☒ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- Data element testing—Cognitive testing
  - 18 respondents were interviewed across 2 rounds of testing.
- Score-level testing
  - Face validity testing, including the method, was reported in the 2011 submission. Review by NCQA's Committee on Performance Measurement (CPM) included identifying measures that meet criteria, definition and testing, public comment, analysis of first-year data collection, determination of suitability for public reporting, and recurrent review of measures.
  - The developers conducted construct validity testing by looking at whether the rates reported in this measure are correlated to each other. They hypothesized that since they assess different aspects of tobacco cessation, that organizations which perform well on one rate should perform well on other rates. Developers examined Pearson correlations to test the strength of the associations.
    - Medicare plans were not included in the analysis since they only report one rate.

- NOTE: An underlying assumption regarding instrument-based performance measures is that they are based on instruments that have been shown to be both reliable and valid. Recently, NQF has operationalized this assumption by requiring demonstration of the reliability and validity of the underlying survey (i.e., data element reliability and validity.)

Validity testing results:

- Cognitive Testing Results
  - Four survey items were tested which corresponded to current smoking status and each of the three rates in the measure.
  - Testing was performed to address issues raised about relevance to guidelines, issues with response bias, and clarity of language.
- Face Validity Testing Results
  - In the 2011 submission, the developer noted that NCQA's Committee on Performance Measurement (CPM) recommended the measure for public reporting (10 supported, 1 opposed, 1 abstained).
- Construct Validity Testing Results
  - Pearson correlations measure the degree of association between two quantitative variables. For the social sciences, scores of 0.37 or larger are considered to have a "large" correlation effect. (Medium effect is 0.24 – 0.36 and small effect is 0.10 – 0.23.)
  - The results indicate that the tobacco cessation rates are significantly correlated with each other in the direction that was hypothesized.

Results of Pearson correlation coefficient analyses (commercial health plans)

- Advising Smokers to Quit and Discussing Cessation Medications rates
  - Pearson correlation coefficient: 0.82
  - P-value: <.0001
- Advising Smokers to Quit and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.77
  - P-value: <.0001
- Discussing Cessation Medications and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.85
  - P-value: <.0001

Results of Pearson correlation coefficient analyses (Medicaid health plans)

- Advising Smokers to Quit and Discussing Cessation Medications rates
  - Pearson correlation coefficient: 0.74
  - P-value: <.0001
- Advising Smokers to Quit and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.68
  - P-value: <.0001
- Discussing Cessation Medications and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.84
  - P-value: <.0001

Questions for the Committee:

- *Does the Committee think there is a need to re-discuss [and re-vote] on validity?*
- *Does the Committee agree with the approach for performing construct validity testing?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

None

Questions for the Committee:

- *Are there any relevant exclusions that should be included here?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

NCQA calculates an inter-quartile range (IQR) for each indicator, which 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, one plan in the 25th percentile and another plan in the 75th percentile of performance. 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 .05, then the two plans' performance is significantly different from each other. The p-value for commercial and Medicaid plans was <0.001 for all three rates, except for the commercial and Medicare Advising Smokers to Quit rates (p-value <0.01).

2016 variation in performance across health plans

Product Line	Rate	Avg. EP	Avg.	Std. Dev.	25th	75th	IQR	p-value
Commercial	Advising Smokers to Quit	132	75%	7%	70%	79%	9%	0.009
	Discussing Cessation Medications	132	48%	8%	43%	52%	9%	<0.001
	Discussing Cessation Strategies	131	44%	9%	38%	50%	12%	<0.001
Medicaid	Advising Smokers to Quit	257	76%	6%	73%	79%	6%	<0.001
	Discussing Cessation Medications	256	48%	8%	43%	54%	11%	<0.001
	Discussing Cessation Strategies	256	43%	7%	39%	48%	9%	<0.001
Medicare	Advising Smokers to Quit	55	86%	6%	82%	90%	8%	0.006

EP: eligible population, the average denominator size across plans submitting to HEDIS

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

Question for the Committee:

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

2b6. Comparability of data sources/methods:

Not needed

2b7. Missing Data

The developer does not provide any information about missing data (e.g., item non-response in survey).

Guidance from the Validity Algorithm

Specifications are consistent with evidence (Box 1)→potential threats mostly assessed (no information about missing data) (Box 2) →empirical reliability testing (Box 3)→score level testing (Box 6)→method sufficient (rates are compared to each other, not a different performance measure) (Box 7) →Moderate certainty (Box 8b) →Moderate

The highest possible rating is HIGH.

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

Committee pre-evaluation comments

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

2a.1 & 2b.1 Specifications: Reliability Specifications

Comments:

**Uses the various CAHPS surveys and sample sizes are calculated to ensure large enough sample to achieve 411 completed surveys per health plan. Calculation algorithm is provided. No exclusions. This reviewer does not think there is a need to re-review ad re-vote on reliability.

**Defined well.

**Initially, in the Brief Measure information, the Denominator is specified as "Patients 18 years and older who responded to the CAHPS survey and indicated that they were current smokers or tobacco users during the measurement year or in the last 6 months for Medicaid and Medicare". However, under the reliability specification, Denominator details are provided for Medicare, Medicaid, and commercial plans. I assume this discrepancy is an error.

**Data elements are clearly defined.

** Specifications seem logical. Not concerned about implementation given this measure's history.

2a.2 Reliability Testing

Comments:

** Updated 2016 data of measure score reliability is provided. To this reviewer, differences in performance can be detected using the reliability score of 0.7.

**Adequate.

**The information provided updating the measure score reliability and construct validity testing is sufficient.

**Developer assessed measure score reliability using data from all health plans that submitted HEDIS data to NCQA for this measure and had a valid rate in 2015/2016. The developer provides data on the denominator sizes for reliability testing, including the number of plans and the median eligible number of patients per plan. The developer provides a detailed explanation of the reliability testing using a beta binomial approach that assesses the signal to noise ration. The results demonstrate sufficient reliability in 2016.
The developer does not correlate responses with prescriptions filled.

** Reliability testing seems adequate. I do not see a need for a revote.

2b.1 Validity Specifications

Comments:

** Specifications are consistent with the evidence.

**Valid--one could quibble about the advantages of additive interventions in NNT.

**Validity was tested using several methodologies: face, construct, and cognitive.

** Specifications seem reasonable given that this is a patient reported measure (and more clinical detail might not be feasible to collect)

2b.2 Validity Testing

Comments:

** Face and empirical validity were tested. New 2016 results of construct validity are provided and additional information about 2008 cognitive testing. For this reviewer, no need to re-discuss and re-vote o validity specifications.

**Adequate.

**Validity was tested using several methodologies: cognitive, face, and construct.

Cognitive Testing Results Four survey items were tested which corresponded to current smoking status and each of the three rates in the measure.

Testing was performed to address issues raised about relevance to guidelines, issues with response bias, and clarity of language. Face Validity Testing Results.

In the 2011 submission, the developer noted that NCQA's Committee on Performance Measurement (CPM) recommended the measure for public reporting (10 supported, 1 opposed, 1 abstained).

Construct Validity Testing Results
The results indicate that the tobacco cessation rates are significantly correlated with each other in the direction that was hypothesized.

** Validity testing seems adequate. I do not see a need for a revote. I thought the approach to construct validity testing was reasonable.

2b.3.-2b7. Testing (Related to Potential Threats to Validity)

Comments:

** No exclusions; no risk adjustment. Meaningful differences calculation suggests that measure can distinguish guality differences.

**Not really.

**The measure depends on recall bias. NCQA does not provide information related to missing data. NCQA does not provide information related to missing data or whether certain categories of respondents are less likely to respond to the CAPHS survey.

NCQA calculates an inter-quartile range (IQR) for each indicator, which 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, one plan in the 25th percentile and another plan in the 75th percentile of performance. The results were statistically significant.

** The IQRs and even more the differences between the 10^{th} and 90^{th} percentiles suggest that there are still meaningful differences in quality

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 required data elements for this measure are collected from a patient-reported survey.
- The patient/family reported information may be obtained via electronic or paper sources.
- Commercial use of a measure requires the prior written consent of NCQA; use by health care physicians in connection with their own practices is not considered commercial use.
- The developer did not report any implementation challenges.

Questions for the Committee:

- *Does recall bias affect the feasibility of this measure?*
- *What is the extent and burden of collecting paper-based information?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

Comments:

**Data elements are collected from patient-reported survey which can be obtained either electronically or on paper; for commercial health plans, NCQA requires written consent. Burden of collecting this information on paper may be high enough to suggest that feasibility is moderate.

**it has been done and date are regularly in EHR.

**NCQA uses several mechanisms to solicit feedback from plans that participate in HEDIS reporting, including a Policy Clarification Support System and a HEDIS Users' Group. The Policy Clarification Support System allows NCQA to collect "real-time" feedback from measure users; through this system, NCQA receives thousands of inquiries each year on over 100 measures. The HEDIS Users' Group has 195 members for 2017; participation includes four conferences presented by NCQA to address key HEDIS implementation issues. NCQA has not heard about difficulties implementing this measure.

** This survey is widely used and clearly feasible.
While, ideally QMs would be collected from clinical data wherever possible, this would exclude patient perspective and be undesirable in this case.

While patient recall could underestimate system performance, on the other hand, provider advice is not useful if not recalled and so this does measure something important about quality of care and might be preferable to provider check boxes.

Criterion 4: Usability and Use

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:

- This measure is included in the Medicaid Adult Core Set and the CMS Quality Rating System (QRS).
- This measure is used by NCQA for scoring in accreditation of commercial, Medicare Advantage, and Medicaid health plans.
    - In 2012, a total of 336 commercial health plans covering 87 million lives, 170 Medicare Advantage health plans covering 7.1 million Medicare beneficiaries, and 77 Medicaid health plans covering 9.1 million lives were accredited using this measure (among others).
- This measure is also reported and used in the following:
    - State of Health Care Annual Report (nationally and by geographic region)
    - Quality Compass
- This measure is used to calculate health plan rankings which are reported in Consumer Reports and on the NCQA website.

Improvement results:

- As noted above, from 2014-2016, all three rates have remained relatively stable across all health plans.
- Developer does not provide reasons for lack of significant change.

Unexpected findings (positive or negative) during implementation: None reported.

Potential harms: None reported.

Vetting of the measure:

- The developer notes they use several methods to obtain input, including several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer notes that health plans that report HEDIS calculate their rates and so know their performance when submitting to NCQA, and that NCQA publicly reports rates across all plans so that the plans can understand their relative performance.
- While the developers provide technical assistance for calculating/implementing the measure, It is not clear whether the developers provide specific technical assistance with interpreting the results.

Feedback:

- In 2016, the MAP Medicaid Task Force again supported the measure's continued use in the Medicaid Adult Core Set.
- No significant barriers to implementing this measure have been reported by health plans.
- Feedback received to date has not required modification to this measure.

***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

4. Usability and Use:

Comments:

**Measure is currently in use in Medicaid, NCQA for accreditation, and in Medicare Advantage. Used by NCQA and Consumer Reports to rank health plans. Rates have remained stable across a number of years, suggesting there is more work to do to provide specific technical assistance to plans in how to understand their results and improve implementation. This issue may reduce the usability for quality improvement from high to moderate.

**Already in use. Well-accepted.

**Can this measure be utilized in BH settings? A very high rate of tobacco use is found among individuals with SUD and Mental illness. And yet the specifications for this measure seem restricted to physician offices and health clinics.

**This measure is being publically reported. This measure is used in accreditation of health plans, reported on Quality Compass, and are used to calculate health plan rankings. Rates have remained unchanged over the past 3 years.

** The measure is widely used. While recent improvement would be preferable, lack of improvement is probably a sign that more effort is needed to move the needle rather than being a case for less measurement.

Criterion 5: Related and Competing Measures

Related or competing measures

0028/3225/3185 : Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention
1654 (TOB-2): Tobacco Use Treatment Provided or Offered
1656 (TOB-3): Tobacco Use Treatment Provided or Offered at Discharge
2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
2803 : Tobacco Use and Help with Quitting Among Adolescents

Harmonization

- Measures 1654 and 1656 are hospital-level measures aimed at offering/providing counseling and cessation interventions
- Measure 2600 focuses on specific populations (SMI, AOD) at the health plan level
- Measure 0028/3225 and 2803 respectively look at screening and cessation interventions in adults and adolescents (respectively) at the clinician group/practice level.
- These measures seem to be mostly harmonized in terms of their definitions, but potential for further harmonization may be discussed at the in-person meeting.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☒ **Yes** ☐ **No**

Pre-meeting public and member comments

- No comments received.

Measure Number (*if previously endorsed*)**:** 0027

Measure Title: Medical Assistance with Smoking and Tobacco Use Cessation

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: 12/2/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*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: Medical Assistance With Smoking and Tobacco Use Cessation

☐ 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.12 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.

2016 Update:

Health care provider screens adults for tobacco use >>> Identifies adults who are using tobacco >>> Advises adults who use tobacco to stop using tobacco and discusses cessation medications and strategies >>> Reduction in tobacco use >>> Improved health >>> Improved health outcomes (including mortality)

Prior submission: This measure is based on a US Preventive Services Task Force guideline. The USPSTF recommends screening and counseling for smoking cessation (A Recommendation). The USPSTF evaluates the effect of a screening or counseling intervention's relationship to health outcomes as part of its analytic framework.

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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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**	U.S. Preventive Services Task Force (USPSTF, 2015): Title: Behavioral and Pharmacotherapy Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Women: U.S. Preventive Services Task Force Recommendation Statement Author: Albert L. Siu, MD, MSPH for the U.S. Preventive Services Task Force Date: October 20, 2015 Citation: Siu, A. L. (2015). Behavioral and Pharmacotherapy Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Women: U.S. Preventive Services Task Force Recommendation Statement. *Annals of Internal Medicine*, 163(8), 622-635. URL: Guidelines available from: http://annals.org/aim/article/2443060/behavioral-pharmacotherapy-interventions-tobacco-smoking-cessation-adults-including-pregnant-women U.S. Preventive Services Task Force (USPSTF, 2009): Title: U.S. Preventive Services Task Force (USPSTF). Counseling and interventions to prevent tobacco use and tobacco-caused disease in adults and pregnant women: U.S. Preventive Services Task Force reaffirmation recommendation statement Author: U.S. Preventive Services Task Force (USPSTF) Date: April 21, 2009 Citation: U.S. Preventive Services Task Force (USPSTF). Counseling and interventions to prevent tobacco use and tobacco-caused disease in adults and pregnant women: U.S. Preventive Services Task Force reaffirmation recommendation statement. Ann Intern Med 2009 Apr 21;150(8):551-5. URL: http://www.guideline.gov/syntheses/synthesis.aspx?id=16422&search=smoking+cessation Institute for Clinical Systems Improvement (ICSI, 2004): Title: Tobacco use prevention and cessation for adults and mature adolescents Author: Institute for Clinical Systems Improvement (ICSI)

	Date: June 24, 2004 Citation: Institute for Clinical Systems Improvement (ICSI). Tobacco use prevention and cessation for adults and mature adolescents. Bloomington (MN): Institute for Clinical Systems Improvement (ICSI); 2004 Jun. 24. p. URL: http://www.guideline.gov/syntheses/synthesis.aspx?id=16422&search=smoking+cessation Veterans' Affairs/Department of Defense (VA/DoD, 2004): Title: VA/DoD clinical practice guideline for the management of tobacco use Author: Veterans Administration/Department of Defense Date: June, 2004 Citation: Veterans Administration, Department of Defense. VA/DoD clinical practice guideline for the management of tobacco use. Washington (DC): Department of Veteran Affairs; 2004 Jun. 81 p. URL: http://www.guideline.gov/syntheses/synthesis.aspx?id=16422&search=smoking+cessation Public Health Service (PHS, 2008): Title: Treating tobacco use and dependence: 2008 update Author: Public Health Service Date: May, 2008 Citation: Public Health Service (PHS). Treating tobacco use and dependence: 2008 update. Rockville (MD): 2008 May. 257 p. URL: http://www.guideline.gov/syntheses/synthesis.aspx?id=16422&search=smoking+cessation
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.	U.S. Preventive Services Task Force (USPSTF, 2015): The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and U.S. Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. (Grade A recommendation) The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. (Grade A recommendation) The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of pharmacotherapy interventions for tobacco cessation in pregnant women. (I statement) The USPSTF concludes that the current evidence is insufficient to recommend electronic nicotine delivery systems (ENDS) for tobacco cessation in adults, including pregnant women. The USPSTF recommends that clinicians direct patients who smoke tobacco to other cessation interventions with established effectiveness and safety (previously stated). (I statement) United States Preventive Services Task Force (USPSTF, 2009): The USPSTF guideline strongly recommends that clinicians screen all adult for tobacco use and provide tobacco cessation interventions for those who use tobacco products. The USPSTF found good evidence that brief smoking cessation interventions, including

	screening, brief behavioral counseling (less than 3 minutes), and pharmacotherapy delivered in primary care settings, are effective in increasing the proportion of smokers who successfully quit smoking and remain abstinent after 1 year (USPSTF, 2003). Institute for Clinical Systems Improvement (ICSI, 2004): The ICSI Tobacco Use Prevention and Cessation for Adults and Mature Adolescents cites tobacco use as the single most preventable cause of disease and death in American society. The guideline recommends that clinicians establish tobacco use for all patients and reassess users are every clinic visit. Assessment of interest in quitting and timing of that interest should be done after the main reasons for the visit have been addressed, and should precede any advice about quitting. This allows a 1 to 3 minute tobacco discussion accommodating both the user's needs and the provider's time limits (ICSI, 2004). Veterans' Affairs/Department of Defense (VA/DoD, 2004): The VA/DoD's Clinical Practice Guideline for the Management of Tobacco Use recommends that any person (age greater than 12 years) who is eligible for care in the Veterans Health Administration (VHA) or the Department of Defense (DoD) health care delivery system should be screened for tobacco use and should be asked about tobacco use at most visits. Tobacco users should be advised to quit and assessed for willingness to quit at every visit. All tobacco users who are willing to quit should be offered an effective tobacco cessation intervention, including: pharmacotherapy, counseling, and follow-up. Tobacco users attempting to quit should be prescribed one or more effective first-line pharmacotherapies for tobacco use cessation. The guideline also cites strong evidence that minimal counseling (lasting less than three minutes) increases overall tobacco abstinence rates. Public Health Service (PHS, 2008): The Public Health Service Clinical Practice Guideline recommends that clinicians engage in a number of activities to aid tobacco users in quitting, which includes: • Implement an officewide system that ensures that, for EVERY patient at EVERY clinic visit, tobacco-use status is queried and documented (repeated assessment is not necessary in the case of the adult who has never used tobacco or has not used tobacco for many years, and for whom this information is clearly documented in the medical record). • In a clear, strong, and personalized manner, urge every tobacco user to quit. • As every tobacco user if he or she is willing to make a quit attempt at this time (e.g., within the next 30 days). • Provide practical counseling (problem solving/training). • Recommend the use of approved pharmacotherapy, except in special circumstances. • Provide supplementary materials.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	2016 Update: Please see grades for the 2015 USPSTF recommendations. The measure is based on multiple guidelines graded A and I. Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Grade I: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.

	Prior Submission: No grade was provided.
Provide all other grades and definitions from the evidence grading system	2016 Update: N/A. Prior Submission: N/A.
Grade assigned to the **recommendation** with definition of the grade	2016 Update: The measure is based on multiple guidelines graded A and I. Grade A: The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Grade I: The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined. Prior Submission: Grade A for the USPSTF 2009 recommendation.
Provide all other grades and definitions from the recommendation grading system	2016 Update: Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. Grade C: The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small. Grade D: The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	2016 Update: Please see the USPSTF Final Evidence Review for quantity and quality of studies used in the body of evidence. Located at https://www.ncbi.nlm.nih.gov/books/NBK321744/ Below, we provide a high-level summary of the USPSTF's quantity and quality of studies. Quantity USPSTF included a total of 54 systematic reviews that met eligibility criteria in its review for this update. Of the 54 reviews included, 22 served as the basis for USPSTF's primary findings. The remaining 32 reviews were listed in descriptive tables. Quality Of the 54 included reviews, 43 addressed tobacco cessation interventions for the adult population with the majority of study designs being RCTs. Of the 43 reviews, nine reviews addressed the effectiveness and/or adverse events related to pharmacotherapy (nicotine replacement therapy (NRT), bupropion hydrochloride sustained release (bupropion SR), and/or varenicline) among the adult population. One review addressed combined pharmacotherapy and behavioral interventions. Twenty-six reviews addressed behavioral tobacco cessation treatments among the adult population. Seven reviews focused on

	specific subpoulations within the general adult population and included behavioral and/or pharmacotherapy interventions. Of the 54 eligible reviews identified, eight were included that evaluated smoking cessation interventions among pregnant women. The majority of study designs included RCTs and quasi-RCTs with few involving cluster-randomized trials, randomized cross-over trials, and prospective cohorts in addition. Of these eight reviews, three reviewed both pharmacotherapy and behavioral interventions, two assessed pharmacotherapy and three assessed only behavioral interventions. Prior Submission: Quantity: The measure is based on a USPSTF guideline that is based on a comprehensive meta-analysis (see USPSTF report for full number of studies) Quality: High
Estimates of benefit and consistency across studies	2016 Update: Please see the USPSTF Final Evidence Review for estimates of benefit and consistency across studies. Located at https://www.ncbi.nlm.nih.gov/books/NBK321744/ Below, we provide a high-level summary of the USPSTF's benefit and consistency across studies. Benefit Of the included pharmacotherapy intervention reviews, there were no existing systematic reviews that assessed pharmacotherapy interventions among adults that reported the effects of interventions on mortality, morbidity, or other health outcomes. Reviews concluded that any for of NRT was beneficial to increasing the rate of smoking cessation, effect of bupropion SR were similar regardless of treatment or recruitment setting, and varenicline reviews provided statistically significant benefits of 1.35 mg daily dose and 0.5 mg twice daily doses, compared with placebo Of the review that assessed the effect of combining pharmacotherapy and behavioral support for smoking cessation among adults, there was a statistically significant benefit of combined pharmacotherapy and behavioral interventions versus control on smoking cessation at 6 months followup or longer. Of the reviews that evaluated the effects of behavioral tobacco cessation among the general adult population, there was considerable overlap in the included studies within groupings (i.e., within the reviews on behavioral support and counseling) and between intervention categories (i.e., behavioral support and counseling and telephone counseling). Of the included reviews that assessed specific subpopulations, the review concluded that more intensive interventions and interventions with combined approaches (pharmacotherapy and followup counseling) achieve the best outcomes. Of the included reviews that assessed pregnant women, the reviews concluded that the impacts on infant health outcomes with NRT were sparse, somewhat mixed, but generally

	favoring no harm or slight benefit and that there was evidence of statistically significant infant health benefits from behavioral interventions. The reviews found there was no evidence of adverse events related to behavioral interventions among pregnant women. Consistency The 32 reviews listed in the descriptive tables were consistent in terms of significance and magnitude of effects in relation to primary reviews. Of the reviews that assessed the effect of pharmacotherapy support for smoking cessation among adults, the USPSTF determined there was a positive net benefit of using NRTs, that bupropion SR reviews were consistent whether group-based or individual-based behavioral therapy, and varenicline reviews were consistent with the time frame tested. There was a single review that assessed the effect of combining pharmacotherapy and behavioral support for smoking cessation among adults. It was consistent with other studies in that interventions were overall positive when compared to placebo. Of the reviews that evaluated the effects of behavioral tobacco cessation among the general adult population, the reviews were widely varied as they were subcatagorized into nine subgroupings. Of the included reviews that assessed specific subpopulations, the reviews were widely varied and included multiple review catwegories; one review concentrated on smokeless tobacco users, four reviews focused on cessation interventions for racial and ethnic minority groups, one only included results for young adults and one focused on interventions among older adult smokers. None of these reviews were considered part of the primary review. Of the included reviews that assessed pregnant women, there was considerably more evidence available on the effects of behavioral interventions during pregnancy than for pharmacotherapies. Prior Submission: Consistent. The USPSTF determined there was a positive net benefit.
What harms were identified?	2016 Update: For a complete summary of the identified harms, please see the USPSTF Final Evidence Review. Located at https://www.ncbi.nlm.nih.gov/books/NBK321744/ Below, we provide a high-level summary of the USPSTF's review of the evidence about harms. General Adult Population NRT users were found to experience minimal harm related to cardiovascular (CV) adverse events, typically low-risk events like tachycardia. Post-marketing research of bupropion sustained release (SR) and varenicline as smoking cessation aids raised concerns regarding patient safety related to neuropsychiatric outcomes (including suicidal ideation and attempts) for bupropion SR as well as serious CV events for varenicline. FDA boxed warnings have been placed on bupropion SR used for

	smoking cessation for possible serious neuropsychiatric adverse events and on varenicline for neuropsychiatric adverse events. The FDA issued a warning for varenicline and its risk of CV adverse events. Continuing research is being conducted for these medications to assess safety. USPSTF found limited evidence of harms related to behavioral interventions used for smoking cessation. Pregnant Women USPSTF found no studies evaluating bupropion SR and varenicline pharmacotherapy harms for pregnant women. Most studies with pregnant women using NRT reported no harm or slight benefit on infant health outcomes. However, in one trial the potential of cesarean section was higher among pregnant women assigned to NRT when compared to placebo group. USPSTF acknowledges that there are few NRT trials for pregnant women and most inconsistently reported adverse events. Electronic nicotine delivery systems (ENDS) have not been approved as a cessation intervention by the FDA and multiple studies are ongoing to effectively assess the harms of ENDS. Prior Submission: N/A.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	2016 Update: We have not identified any new studies conducted since the systematic review. Prior 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.

N/A

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

N/A

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

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

0027_MSC_Evidence_Form_2016-636179402077250670.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

Tobacco smoking is the leading cause of preventable disease and death in the United States, resulting in approximately 480,000 premature deaths and more than $300 billion in direct health care expenditures and productivity losses each year (HHS, 2014). Premature deaths due to smoking, including deaths from lung cancer, pulmonary diseases, coronary heart disease, pregnancy concerns, and residential fires, numbered over 20 million between 1965 and 2014 (HHS, 2014). Although the consumption of cigarettes continues to decline (with a decrease from 20.9 percent in 2005 to 16.8 percent in 2014 (Centers for Disease Control and Prevention, 2015)), the use of electronic cigarettes, or e-cigarettes, has more than doubled between 2011 and 2012, especially among adolescents (HHS, 2014).

The strongest evidence on increasing smoking cessation comes from studies involving physician or nurse's advice, tailored self-help materials, or telephone counseling (Siu, 2015). For example, interventions that involve physician or nurse advice are associated with smoking abstinence at six months or more after the intervention (8.0 percent for physicians and 13.3 percent for nurses) compared to no advice or usual care (4.8 percent for physicians and 11.3 percent for nurses). Study participants who receive tailored self-help materials are more likely to cease smoking at six months or more when compared to study participants who did not receive self-help materials (7.1 percent vs. 5.8 percent). The U.S. Prevention Services Task Force (USPSTF) has found evidence that smoking cessation decreases the risk for heart disease, lung disease, and stroke through a review of published literature. The USPSTF also highlights evidence that smoking and tobacco use cessation interventions (including counseling sessions and pharmacotherapy) are effective in increasing the proportion of patients who remain tobacco-free for at least 6 months to 1 year depending on length of intervention (Siu, 2015).

Citations:

Centers for Disease Control and Prevention. (2015). Current Cigarette Smoking Among Adults—United States, 2005–2014. Morbidity and Mortality Weekly Report;64(44):1233–40. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6444a2.htm

Siu, A. L. (2015). Behavioral and Pharmacotherapy Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Women: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine, 163(8), 622-635. Retrieved from: http://annals.org/aim/article/2443060/behavioral-pharmacotherapy-interventions-tobacco-smoking-cessation-adults-including-pregnant-women

U.S. Department of Health and Human Services (HHS). (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Retrieved from https://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b) 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 is summarized at the health plan level and summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. Data is stratified by year and product line (i.e. commercial, Medicaid, Medicare).

The following data demonstrate the variation in performance for all three rates across health plans. The difference in performance between plans in the 10th and 90th percentiles varies from 12 to 27 points across the three rates and product lines. This difference is greater for the Discussing Cessation Medications and Discussing Cessation Strategies rates; in 2016, for the Medicaid product line, the difference was 21 and 18 percentage points, respectively; for the commercial product line, the difference was 20 and 24 percentage points, respectively. In 2016, for the Advising Smokers to Quit rate, the difference was 15 percentage points for the Medicare product line, 17 percentage points for the commercial product line, and 14 percentage points for the Medicaid product line. These gaps in performance underscore the ongoing opportunity for improvement.

Advising Smokers to Quit
Medicare Rate

YEAR	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX	Interquartile Range
2014	84%	7%	61%	75%	81%	85%	89%	92%	100%	8%
2015	86%	6%	69%	78%	82%	86%	90%	93%	98%	8%
2016	86%	6%	69%	78%	82%	86%	90%	93%	98%	8%

Advising Smokers to Quit Rolling Average
Commercial Rate

YEAR	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX	Interquartile Range
2014	75%	8%	57%	66%	69%	74%	80%	85%	92%	11%
2015	75%	7%	60%	64%	70%	76%	79%	83%	92%	9%
2016	75%	7%	59%	66%	70%	75%	79%	83%	92%	9%

Advising Smokers to Quit Rolling Average
Medicaid Rate
Commercial Rate

YEAR	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX	Interquartile Range
2014	76%	5%	54%	69%	74%	77%	79%	81%	87%	5%
2015	76%	6%	54%	68%	74%	77%	79%	82%	86%	5%
2016	76%	6%	60%	68%	73%	77%	79%	82%	91%	6%

Discussing Cessation Medications
Commercial Rate

YEAR	MEAN	ST DEV	MIN	10TH	25TH	50TH	75TH	90TH	MAX	Interquartile Range
2014	49%	9%	29%	37%	42%	49%	55%	61%	73%	13%
2015	49%	8%	32%	38%	43%	50%	53%	58%	71%	10%

2016 | 48% | 8% | 31% | 41% | 43% | 48% | 52% | 61% | 74% | 9%

Discussing Cessation Medications
Medicaid Rate
YEAR| MEAN | ST DEV | MIN | 10TH | 25TH | 50TH | 75TH | 90TH | MAX | Interquartile Range
2014 | 47% | 8% | 28% | 38% | 41% | 46% | 52% | 57% | 67% | 11%
2015 | 47% | 8% | 19% | 36% | 42% | 47% | 52% | 57% | 65% | 10%
2016 | 48% | 8% | 25% | 37% | 43% | 48% | 54% | 58% | 66% | 11%

Discussing Cessation Strategies
Commercial Rate
YEAR| MEAN | ST DEV | MIN | 10TH | 25TH | 50TH | 75TH | 90TH | MAX | Interquartile Range
2014 | 43% | 10% | 19% | 30% | 36% | 42% | 49% | 57% | 67% | 13%
2015 | 44% | 9% | 26% | 32% | 38% | 43% | 50% | 56% | 65% | 12%
2016 | 44% | 9% | 28% | 34% | 38% | 42% | 50% | 58% | 65% | 12%

Discussing Cessation Strategies
Medicaid Rate
YEAR| MEAN | ST DEV | MIN | 10TH | 25TH | 50TH | 75TH | 90TH | MAX | Interquartile Range
2014 | 42% | 7% | 26% | 34% | 38% | 42% | 45% | 51% | 59% | 7%
2015 | 42% | 7% | 23% | 34% | 38% | 43% | 48% | 51% | 56% | 10%
2016 | 43% | 7% | 27% | 34% | 39% | 44% | 48% | 52% | 61% | 9%

In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare 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.

Advising Smokers to Quit
Medicare Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 316 | 57
2015 | 221 | 55
2016 | 238 | 55

Advising Smokers to Quit Rolling Average
Commercial Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 127 | 130
2015 | 83 | 131
2016 | 59 | 132

Advising Smokers to Quit Rolling Average
Medicaid Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 137 | 275
2015 | 139 | 274
2016 | 159 | 257

Discussing Cessation Medications
Commercial Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 126 | 130
2015 | 83 | 130

2016 | 58 | 132

Discussing Cessation Medications
Medicaid Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 137 | 273
2015 | 138 | 274
2016 | 159 | 256

Discussing Cessation Strategies
Commercial Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 123 | 130
2015 | 82 | 130
2016 | 58 | 131

Discussing Cessation Strategies
Medicaid Rate
YEAR | N Plans | Mean Denominator Size per plan
2014 | 137 | 273
2015 | 139 | 272
2016 | 159 | 256

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 (4b) under Usability and Use.*
NCQA does not currently report performance data stratified by race or ethnicity. While not specified in the measure, results from CAHPS surveys can be stratified by the health plan for demographic variables, such as race/ethnicity collected from the survey.

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
Although HEDIS measures are not stratified by race and ethnicity, researchers have explored disparities in the provision of smoking and tobacco cessation strategies and medications. Significant disparities in tobacco use are still present among certain racial/ethnic populations, and among groups defined by educational level, geographic region, sexual minorities (including the gay, lesbian, bisexual, and transgender community, and individuals with same-sex relationships or attraction), and severe mental illness (HHS, 2014). Men are more likely to be current smokers than women; 18.8 percent of men reported being current smokers in 2014 and 14.8 percent of women. Among racial and ethnic groups, the smoking prevalence is highest among American Indian/Alaska Natives (29.2 percent) and multiracial adults (27.9 percent), and lowest among Asians (9.5 percent). In regards to education, among adults aged =25 years, the prevalence was highest among persons with a General Education Development (GED) certificate (43.0 percent) and lowest among those with a graduate degree (5.4 percent). The occurrence of tobacco use is higher among lesbian, gay, or bisexual adults (23.9 percent) than among straight adults (16.6 percent). By region, Individuals living in the Midwest have the highest prevalence of tobacco use (20.7 percent) compared to those living in the West (13.1 percent). Individuals who do not report a disability or limitation have a lower prevalence of tobacco use (16.1 percent) compared to those who do report having a disability or limitation (23.9 percent) (Centers for Disease Control and Prevention, 2015).

Research studies are limited in information regarding smoking cessation and sexual minorities, severe mental illness, and geographic region. Lower socioeconomic individuals are more likely to discontinue pharmacotherapy interventions and have barriers to completing behavioral methods due to lack of knowledge, cost, and low self-efficacy (Hiscock et al., 2012). Adults >25 in age are more likely to use pharmacological treatments to aid in smoking cessation than their younger counterparts who are more likely to use support from friends or family (Curry et al., 2007). Women are more likely than men to have poorer smoking cessation outcomes due to multiple facts; perceived weight gain during cessation, lower self-efficacy, and anticipated negative tobacco use withdrawal symptoms (McKee et al., 2005).

Citations:

Centers for Disease Control and Prevention. (2015). Current Cigarette Smoking Among Adults—United States, 2005–2014. Morbidity and Mortality Weekly Report;64(44):1233–40. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6444a2.htm

Curry, S. J., Sporer, A. K., Pugach, O., Campbell, R. T., & Emery, S. (2007). Use of Tobacco Cessation Treatments Among Young Adult Smokers: 2005 National Health Interview Survey. American journal of public health, 97(8), 1464-1469.

Hiscock, R., Bauld, L., Amos, A., Fidler, J. A., & Munafò, M. (2012). Socioeconomic Status and Smoking: A Review. Annals of the New York Academy of Sciences, 1248(1), 107-123.

McKee, S. A., O´Malley, S. S., Salovey, P., Krishnan-Sarin, S., & Mazure, C. M. (2005). Perceived Risks and Benefits of Smoking Cessation: Gender-Specific Predictors of Motivation and Treatment Outcome. Addictive behaviors, 30(3), 423-435.

U.S. Department of Health and Human Services (HHS). (2014). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Retrieved from https://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf.

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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.)*
NA

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.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.
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).
Advising Smokers and Tobacco Users to Quit:
Patients who indicated that they received advice to quit smoking or using tobacco from their doctor or health provider

Discussing Cessation Medications:
Patients who indicated that their doctor or health provider recommended or discussed smoking or tobacco cessation medications

Discussing Cessation Strategies:
Patients who indicated their doctor or health provider discussed or provided smoking or tobacco cessation methods and strategies other than medication

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
For the commercial product line:
- Advising Smokers and Tobacco Users to Quit:
The number of patients in the denominator who indicated that they received advice to quit smoking or tobacco use from a doctor or other health provider by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q47: “In the last 12 months, how often were you advised to quit smoking or using tobacco by a doctor or other health provider in your plan?”

- Discussing Smoking Cessation Medications:

The number of patients in the denominator who indicated that their doctor or health provider recommended or discussed medication to assist with quitting smoking or using tobacco by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q48: “In the last 12 months, how often was medication recommended or discussed by a doctor or health provider to assist you with quitting smoking or using tobacco? Examples of medication are: nicotine gum, patch, nasal spray, inhaler, or prescription medication.”

- Discussing Cessation Strategies:
The number of patients in the denominator who indicated that their doctor or health provider discussed or provided methods and strategies other than medication to assist with quitting smoking or using tobacco by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q49: “In the last 12 months, how often did your doctor or health provider discuss or provide methods and strategies other than medication to assist you with quitting smoking or using tobacco? Examples of methods and strategies are: telephone helpline, individual or group counseling, or cessation program.”

Response options for all questions:
Never, Sometimes, Usually, Always

For the Medicaid product line:
- Advising Smokers and Tobacco Users to Quit:
The number of patients in the denominator who indicated that they received advice to quit smoking or tobacco use from a doctor or other health provider by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q40: “In the last 6 months, how often were you advised to quit smoking or using tobacco by a doctor or other health provider in your plan?”

- Discussing Smoking Cessation Medications:
The number of patients in the denominator who indicated that their doctor or health provider recommended or discussed medication to assist with quitting smoking or using tobacco by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q41: “In the last 6 months, how often was medication recommended or discussed by a doctor or health provider to assist you with quitting smoking or using tobacco? Examples of medication are: nicotine gum, patch, nasal spray, inhaler, or prescription medication.”

- Discussing Cessation Strategies:
The number of patients in the denominator who indicated that their doctor or health provider discussed or provided methods and strategies other than medication to assist with quitting smoking or using tobacco by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q42: “In the last 6 months, how often did your doctor or health provider discuss or provide methods and strategies other than medication to assist you with quitting smoking or using tobacco? Examples of methods and strategies are: telephone helpline, individual or group counseling, or cessation program.”

Response options for all questions:
Never, Sometimes, Usually, Always

For the Medicare product line:
- Advising Smokers or Tobacco Users to Quit
The number of patients in the denominator who indicated that they received advice to quit smoking or using tobacco from a doctor or other health provider by answering “Sometimes” or “Usually” or “Always” to CAHPS question Q66 : “In the last 6 months, how often were you advised to quit smoking or using tobacco by a doctor or other health provider in your plan?”

Response options for all questions:
Never, Sometimes, Usually, Always, I had no visits in the last 6 months

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients 18 years and older who responded to the CAHPS survey and indicated that they were current smokers or tobacco users during the measurement year or in the last 6 months for Medicaid and Medicare.

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).
In order to be included in the denominator for each rate, patients must answer both the question about current cigarette/tobacco use and the relevant numerator question (eg, for the Advising Smokers and Tobacco Users to Quit rate, patients must answer the question about current cigarette/tobacco use and the question about how often they were advised to quit by a doctor or other health provider).

For the commercial product line:
- Advising Smokers and Tobacco Users to Quit
The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering "Every day" or "Some days" to CAHPS question Q46 and by answering Q47 with any response ("Never" or "Sometimes" or "Usually" or "Always").

Q46: "Do you now smoke cigarettes or use tobacco every day, some days, or not at all?"
Response options for Q46: "Every day", "Some days", "Not at all", "Don´t know"

Q47: "In the last 12 months, how often were you advised to quit smoking or using tobacco by a doctor or other health provider in your plan?"
Response options for Q47: "Never", "Sometimes", "Usually", "Always"

- Discussing Cessation Medications
The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering "Every day" or "Some days" to CAHPS question Q46 and by answering Q48 with any response ("Never" or "Sometimes" or "Usually" or "Always").

Q46: "Do you now smoke cigarettes or use tobacco every day, some days, or not at all?"
Response options for Q46: "Every day", "Some days", "Not at all", "Don´t know"

Q48: "In the last 12 months, how often was medication recommended or discussed by a doctor or health provider to assist you with quitting smoking or using tobacco? Examples of medication are: nicotine gum, patch, nasal spray, inhaler, or prescription medication."
Response options for Q48: "Never" OR "Sometimes" OR "Usually" OR "Always"

- Discussing Cessation Strategies
The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering "Every day" or "Some days" to CAHPS question Q46 and by answering Q49 with any response ("Never" or "Sometimes" or "Usually" or "Always").

Q46: "Do you now smoke cigarettes or use tobacco every day, some days, or not at all?"
Response options for Q46: "Every day", "Some days", "Not at all", "Don´t know"

Q49: "In the last 12 months, how often did your doctor or health provider discuss or provide methods and strategies other than medication to assist you with quitting smoking or using tobacco? Examples of methods and strategies are: telephone helpline, individual or group counseling, or cessation program."
Response options for Q49: "Never", "Sometimes", "Usually", "Always"

For the Medicaid product line:
- Advising Smokers and Tobacco Users to Quit

The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering “Every day” or “Some days” to CAHPS question Q39 and by answering Q40 with any response (“Never” or “Sometimes” or “Usually” or “Always”).

Q39: “Do you now smoke cigarettes or use tobacco every day, some days, or not at all?”
Response options for Q39: “Every day”, “Some days”, “Not at all”, “Don´t know”

Q40: “In the last 6 months, how often were you advised to quit smoking or using tobacco by a doctor or other health provider in your plan?”
Response options for Q40: “Never”, “Sometimes”, “Usually”, “Always”

- Discussing Cessation Medications
The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering “Every day” or “Some days” to CAHPS question Q39 and by answering Q41 with any response (“Never” or “Sometimes” or “Usually” or “Always”).

Q39: “Do you now smoke cigarettes or use tobacco every day, some days, or not at all?”
Response options for Q39: “Every day”, “Some days”, “Not at all”, “Don´t know”

Q41: “In the last 6 months, how often was medication recommended or discussed by a doctor or health provider to assist you with quitting smoking or using tobacco? Examples of medication are: nicotine gum, patch, nasal spray, inhaler, or prescription medication.”
Response options for Q41: “Never”, “Sometimes”, “Usually”, “Always”

- Discussing Cessation Strategies
The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering “Every day” or “Some days” to CAHPS question Q39 and by answering Q42 with any response (“Never” or “Sometimes” or “Usually” or “Always”).

Q39: “Do you now smoke cigarettes or use tobacco every day, some days, or not at all?”
Response options for Q39: “Every day”, “Some days”, “Not at all”, “Don´t know”

Q42: “In the last 6 months, how often did your doctor or health provider discuss or provide methods and strategies other than medication to assist you with quitting smoking or using tobacco? Examples of methods and strategies are: telephone helpline, individual or group counseling, or cessation program.”
Response options for Q42: “Never”, “Sometimes”, “Usually”, “Always”

For the Medicare product line:
- Advising Smokers or Tobacco Users to Quit
The number of patients who responded to the survey and indicated that they were current smokers or tobacco users by answering “Every day” or “Some days” to CAHPS question Q65, had one or more visits during the last 6 months, and by answering Q66 with any response (“Never” or “Sometimes” or “Usually” or “Always”).

Q65: “Do you now smoke cigarettes or use tobacco every day, some days, or not at all?”
Response options for Q65: “Not at all”, “Some days”, “Every day”, “Don’t know”

Q66: “In the last 6 months, how often were you advised to quit smoking or using tobacco by a doctor or other health provider in your plan?”
Response options for Q66: “Never”, “Sometimes”, “Usually”, “Always”, “I had no visits in the last 6 months”

The Medicare results for the Advising Smokers and Tobacco Users to Quit Rate requires a minimum denominator of at least 30 responses.

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.)*
N/A

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
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: Identify the eligible population of commercial, Medicaid and Medicare CAHPS respondents

Step 2: Identify the denominator for each component.

Step 3: Identify the numerator for each component.

Step 4: Calculate the rate as numerator/denominator.

For the commercial and Medicaid product lines, rolling averages are calculated using the formula below.
Rate = (Year 1 Numerator + Year 2 Numerator)/(Year 1 Denominator + Year 2 Denominator)

NCQA calculates a result when the denominator is 100 individuals or more.

If the health plan did not report results in the prior year (Year 1), but reports results for the current year and achieves a denominator of 100 or more, NCQA calculates a rate.

For the Medicare product line, this is collected by the Centers for Medicare & Medicaid Services through the Medicare CAHPS Survey. This is collected on an annual basis.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
A systematic sampling method is used for the CAHPS survey. Required sample sizes are based on the goal of achieving 411 complete and valid surveys. To establish required sample sizes, NCQA evaluates a health plan's prior year's survey results and analyzes the following.

• Survey response rates (mean, median and distribution of response rates)
• The average number of complete and valid surveys achieved
• The percentage of members in the sample who were ineligible
• The percentage of members in the sample who, because of a bad address or telephone number, were unable to be contacted by the survey vendor
• The percentage of members who refused to participate in the survey

For each HEDIS/CAHPS survey administered, the survey vendor draws a random sample of members, employing the required sample size as indicated in Table S-3. In a health plan with fewer eligible members than the required sample size, the sample includes the health plan's entire eligible population. To reduce respondent burden, the survey vendor deduplicates samples so that only one adult member per household is included in the adult sample and only one child member per household is included in the child sample.

Table S-3: Survey Sample Sizes
Survey Type Required Sample Size
Adult commercial 1,100
Adult Medicaid 1,350
Child commercial 900
Child Medicaid 1,650

Proxy Responses: Proxy responses are not permitted for the adult CAHPS survey; the sampled member must complete his or her own survey.

Medicare CAHPS: CMS provide the sample for the Medicare CAHPS 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.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Health plans select one of two standard options for administering the CAHPS surveys:
1. The mail-only methodology, a five-wave mail protocol with three questionnaire mailings and two reminder postcards.
2. The mixed methodology, a four-wave mail protocol (two questionnaires and two reminder postcards) with telephone follow-up of a minimum of three and a maximum of six telephone attempts.

Confidentiality of sampled members must be maintained. Neither NCQA nor the health plan has access to the names of individuals selected for the survey.

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

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
CAHPS Health Plan Survey 5.0H, Adult Version; Medicare CAHPS

http://www.ahrq.gov/cahps/index.html

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

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Health Plan, Integrated Delivery System

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Clinician Office/Clinic, Other
If other: In addition to clinician visits, some respondents may recall contacts with an "other health provider" (the wording used in the survey question), which may include contacts with nurses or health plan staff.

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
0027_MSC_Testing_Form_2016_Updated.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. (Do not remove prior testing information – include date of new information in red.)
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. (Do not remove prior testing information – include date of new information in red.)
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
No - This measure is not risk-adjusted

Measure Number (*if previously endorsed*)**:** 0027
Measure Title: Medical Assistance with Smoking and Tobacco Use Cessation
Date of Submission: 12/2/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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*).
N/A

1.3. What are the dates of the data used in testing? Initial testing of measure score reliability was performed using HEDIS performance measurement 2010 data. For the 2016 update, we assessed measure score reliability using data from all health plans that submitted HEDIS data to NCQA for this measure and had a valid rate in 2015/2016, which used data submitted to NCQA in 2016. NCQA conducted cognitive testing of the questions in 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.26*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

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

2016 update: measure score reliability and construct validity testing: Measure score reliability was calculated for the three rates from the 59 commercial health plans, 159 Medicaid health plans, and 238 Medicare health plans that submitted data on this measure to HEDIS in 2016 (58 commercial plans submitted valid rates for the Discussing Cessation Medications and Discussing Cessation Strategies rates). Construct validity was calculated using the same commercial and Medicaid health plans (construct validity could not be calculated for the Medicare plans as they only submit one rate, Advising Smokers to Quit). Commercial and Medicaid plans are included in these analyses only if their denominator contains at least 100 individuals; Medicare plans must have a minimum of 30 individuals in their denominator to be included. The plans were geographically diverse and varied in size.

Systematic evaluation of face validity: This measure was tested for face validity with two panels of experts. Measurement Advisory Panels and subject matter workgroups provide the clinical and technical knowledge required to develop the measures. The Smoking Measurement Workgroup included nine experts in smoking and tobacco use and included representation from consumers, health plans, health care providers and policy makers. NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the HEDIS measurement set and includes representation from purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 21

members. The CPM is organized and managed by NCQA, and is responsible for advising NCQA staff on the development and maintenance of performance measures. The CPM also meets with the NCQA Board of Directors to recommend measures for inclusion in HEDIS. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement. Additional HEDIS Expert Panels and the Technical Advisory Group provide invaluable assistance by identifying methodological issues and giving feedback on new and existing measures. See *Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development* for names and affiliations of expert panel members.

Initial testing of the CAHPS survey instrument:
There are two different and complementary approaches to assessing the reliability and validity of a questionnaire:
1. Cognitive testing, which bases its assessments on feedback from interviews with people who are asked to react to the survey questions.
2. Psychometric testing, which consists of analyses of data collected using the questionnaire.

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*)

2016 update: measure score reliability and construct validity testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, 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 sample. It includes number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Table 1. Denominator sizes for reliability and construct validity testing

Product type	Rate	Number of plans	Median number of eligible patients per plan
Commercial	Advising Smokers to Quit	59	132
	Discussing Cessation Medications	58	132
	Discussing Cessation Strategies	58	132
Medicaid	Advising Smokers to Quit	159	257
	Discussing Cessation Medications	159	256
	Discussing Cessation Strategies	159	256
Medicare	Advising Smokers to Quit	238	55

Patient sample for cognitive testing of survey questions: A total of 18 respondents were interviewed across two rounds of cognitive testing; age ranged from 26 to 69 years of age. Respondents were recruited for variation in race/ethnicity, level of smoking, and type of insurance—commercial, Medicare, and Medicaid.

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.

The samples are described above.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data

are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

2016 update: Measure performance results are stratified by commercial, Medicaid and Medicare health plans.

_____

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*)

Method for initial measure score reliability testing: In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009) in work produced for the National Committee for Quality Assurance (NCQA).

The following is quoted from the tutorial which focused on provider-level assessment: "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." This approach is also relevant to health plans and other accountable entities.

The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. 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 accountable entity variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Adams' approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures.

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

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*)

2016 update: results for measure score reliability testing:

Table 2. Beta-binomial statistic for each measure rate

	Commercial	Medicaid

Advising Smokers to Quit	0.69	0.75
Discussing Cessation Medications	0.72	0.83
Discussing Cessation Strategies	0.77	0.77

Results for initial measure score reliability testing:
1. Commercial plans 2010:
1.a. Advising Smokers & Tobacco Users to Quit: 0.618960
1.b. Discussing Cessation Medications: 0.469075
1.c. Discussing Cessation Strategies: 0.700878

2. Medicaid 2010:
2.a. Advising Smokers & Tobacco Users to Quit: 0.605218
2.b. Discussing Cessation Medications: 0.851239
2.c. Discussing Cessation Strategies: 0.790183

3. Medicare 2010
3.a. Advising Smokers & Tobacco Users to Quit Only: 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?*)

2016 update: interpretation of results for measure score reliability testing: Generally, a reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Beta binomial testing for this measure suggests that the three indicators within this measure have demonstrated good reliability.

2b2. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

2016 update: method for assessing construct validity: We tested for construct validity by exploring whether the rates within this measure were correlated with each other, since they assess different aspects of tobacco use cessation. We hypothesized that organizations that perform well on one rate should perform well on the other rates. 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 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variables 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.

For this measure, we specifically hypothesized:

1) The Advising Smokers to Quit rate will be positively correlated with the Discussing Cessation Medications rate.
2) The Advising Smokers to Quit rate will be positively correlated with the Discussing Cessation Strategies rate.
3) The Discussing Cessation Medications rate will be positively correlated with the Discussing Cessation Strategies rate.

To note, Medicare plans are not included in these analyses, because they only report the Advising Smokers to Quit rate.

Method for initial systematic assessment of face validity:
NCQA identified and refined measure management into a standardized process called the HEDIS measure life cycle.

*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 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 MAPs, the TAG, the HEDIS Policy Panel and various other panels.

*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.Ensure funding throughout measure testing
2.Prepare a detailed conceptual and operational work-up that includes a testing proposal
3.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 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 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 Quality Compass? or in accreditation scoring.

The first-year distinction guarantees that a measure can be efficiently 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 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. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments contribute to measure evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, a third of the measurement set is researched for changes in clinical guidelines or 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 next year's HEDIS Volume 2.

What makes a measure "Desirable"?

Whether considering the value of a new measure or the continuing worth of an existing one, we must define what makes a measure useful. HEDIS measures encourage improvement. The defining question for all performance measurement—"Where can measurement make a difference?"—can be answered only after considering many factors. NCQA has established three areas of desirable characteristics for HEDIS measures, discussed below.

1. Relevance: Measures should address features that apply to purchasers or consumers, or which will stimulate internal efforts toward quality improvement. More specifically, relevance includes the following attributes.

Meaningful: What is the significance of the measure to the different groups concerned with health care? Is the measure easily interpreted? Are the results meaningful to target audiences?
Measures should be meaningful to at least one HEDIS audience (e.g., individual consumers, purchasers or health care systems). Decision makers should be able to understand a measure's clinical and economic significance.

Important to health: What is the prevalence and overall impact of the condition in the U.S. population? What significant health care aspects will the measure address?
We should consider the type of measure (e.g., outcome or process), the prevalence of medical condition addressed by the measure and the seriousness of affected health outcomes.

Financially important: What financial implications result from actions evaluated by the measure? Does the measure relate to activities with high financial impact?
Measures should relate to activities that have high financial impact.

Cost effective: What is the cost benefit of implementing the change in the health care system? Does the measure encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness?

Measures should encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness.

Strategically important: What are the policy implications? Does the measure encourage activities that use resources efficiently? Measures should encourage activities that use resources most efficiently to maximize member health.

Controllable: What impact can the organization have on the condition or disease? What impact can the organization have on the measure? Health care systems should be able to improve their performance. For outcome measures, at least one process should be controlled and have an important effect on outcome. For process measures, there should be a strong link between the process and desired outcome.

Variation across systems: Will there be variation across systems? There should be the potential for wide variation across systems.

Potential for improvement: Will organizations be able to improve performance? There should be substantial room for performance improvement.

2. Scientific soundness: Perhaps in no other industry is scientific soundness as important as in health care. Scientific soundness must be a core value of our health care system—a system that has extended and improved the lives of countless individuals.

Clinical evidence: Is there strong evidence to support the measure? Are there published guidelines for the condition? Do the guidelines discuss aspects of the measure? Does evidence document a link between clinical processes and outcomes addressed by the measure? There should be evidence documenting a link between clinical processes and outcomes.

Reproducible: Are results consistent? Measures should produce the same results when repeated in the same population and setting.

Valid: Does the measure make sense? Measures should make sense logically and clinically, and should correlate well with other measures of the same aspects of care.

Accurate: How well does the measure evaluate what is happening? Measures should precisely evaluate what is actually happening.

Risk adjustment: Is it appropriate to stratify the measure by age or another variable? Measure variables should not differ appreciably beyond the health care system's control, or variables should be known and measurable. Risk stratification or a validated model for calculating an adjusted result can be used for measures with confounding variables.

Comparability of data sources: How do different systems affect accuracy, reproducibility and validity? Accuracy, reproducibility and validity should not be affected if different systems use different data sources for a measure.

3. Feasibility:
The goal is not only to include feasible measures, but also to catalyze a process whereby relevant measures can be made feasible.

Precise specifications: Are there clear specifications for data sources and methods for data collection and reporting? Measures should have clear specifications for data sources and methods for data collection and reporting.

Reasonable cost: Does the measure impose a burden on health care systems? Measures should not impose an inappropriate burden on health care systems.

Confidentiality: Does data collection meet accepted standards of member confidentiality?
Data collection should not violate accepted standards of member confidentiality. Logistical feasibility
Are the required data available?

Auditability: Is the measure susceptible to exploitation or "gaming" that would be undetectable in an audit? Measures should not be susceptible to manipulation that would be undetectable in an audit.

Method for initial assessment of CAHPS survey instrument: Cognitive testing provides useful information about respondents' comprehension of the questions, their ability to answer the questions, and the adequacy of the response choices. It also helps identify words that can be used to describe health care providers accurately and consistently across a range of consumers (e.g., commercially insured, Medicaid, fee-for-service, managed care, lower socioeconomic status (SES), middle SES, low literacy, higher literacy) and explores whether key words and concepts work equally well in both English and Spanish.

Field tests and psychometric analyses provide information about the items' reliability and validity. Many existing questionnaires about health care have been tested primarily or exclusively using a psychometric approach, but the CAHPS team views the combination of cognitive and psychometric approaches as essential to producing the best possible survey instrument.

Additional information about cognitive testing of survey questions: The intent of the cognitive testing NQCA conducted in 2008 was to test four smoking cessation survey items from CAHPS 4.0 to address issues raised by health plans and the expert workgroup during the re-evaluation of the Medical Assistance with Smoking Cessation HEDIS measure. The recommended revisions to the questions addressed relevance with current guidelines (at the time), issues with response bias and clarity of survey language.

Cognitive testing was conducted in two rounds. In Round 1, all respondents were tested using the paper and pencil instrument (PAPI). Two versions of the protocol were tested in Round 2, a PAPI version and a telephone instrument version. The PAPI version was administered to the first four respondents and the telephone instrument was administered to the remaining five respondents.

For the PAPI testing, respondents were asked to read each item aloud and to provide a response. Probing was conducted concurrently, that is, administered after respondents answered each question. Interviewers administered the probes contained in the scripted protocol and followed up on other topics that emerged in testing. Respondents were given a $50 incentive to thank them for their time.

For the telephone instrument, the interviewer explained to the respondent that this is a survey that will be conducted by telephone. To simulate a telephone survey environment, the interviewer dialed in via phone to the respondent from another location. The interviewer read each question aloud to the respondent and noted the responses. After completing the survey by phone, the interviewer re-entered the room and administered the protocol retrospectively. For ease of reference, the respondent was able to look at each question and the answer they had provided.

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

2016 update: results from assessment of construct validity: The results indicate that the tobacco cessation rates are significantly correlated with each other in the direction that was hypothesized.

Results of Pearson correlation coefficient analyses (commercial health plans)

- Advising Smokers to Quit and Discussing Cessation Medications rates
  - Pearson correlation coefficient: 0.82
  - P-value: <.0001
- Advising Smokers to Quit and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.77
  - P-value: <.0001
- Discussing Cessation Medications and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.85
  - P-value: <.0001

Results of Pearson correlation coefficient analyses (Medicaid health plans)

- Advising Smokers to Quit and Discussing Cessation Medications rates
  - Pearson correlation coefficient: 0.74
  - P-value: <.0001
- Advising Smokers to Quit and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.68
  - P-value: <.0001
- Discussing Cessation Medications and Discussing Cessation Strategies rates
  - Pearson correlation coefficient: 0.84
  - P-value: <.0001

Results for initial systematic assessment of face validity:

Step 1: The Medical Assistance with Smoking Cessation was reevaluated in 2008. NCQA's Performance Measurement Department and the Smoking Cessation Measurement Workgroup worked together to rename the measure to Medical Assistance with Smoking and Tobacco Use Cessation and review the cognitive testing results of the CAPHS survey.

Step 2: The proposed measure revisions and cognitive testing results were presented to the CPM in 2009. The CAHPS Survey has been deemed valid as a survey instrument. The CPM recommended to send the measure to public comment with a vote of 12 in favor and none opposed.

Step 3: The measure was released for Public Comment in spring 2009. We received and responded to comments on this measure. The CPM recommended moving this measure to first year data collection with a vote of 10 in favor and none opposed.

Step 4: The Medical Assistance with Smoking and Tobacco Use Cessation measure was introduced in HEDIS 2010. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following two years. The Cardiovascular MAP assumed the responsibilities of the Smoking Cessation Measurement Workgroup reviewed the first year results and recommended public reporting. The CPM recommended moving this measure public reporting with a vote of 10 in favor 1 opposed, and 1 abstained.

Results from cognitive testing of survey questions: For the first topic, related to an individual's current smoking status, respondents were asked to answer two questions: one that asked about cigarette use only, and one that asked about cigarette or tobacco use. Respondents were able to interpret the term "tobacco", had no difficulty with the question, and preferred the question that used both smoking and tobacco to the question that used only smoking.

For the second topic, related to an individual's experience with receiving advice to quit smoking by a doctor or other health provider, respondents were asked to answer two questions: one that used response options about the number of visits (eg, "1 visit", "2 to 4 visits", etc.) and one that used open quantifier response options (eg, "never", "once",

"sometimes", etc.). Respondents were generally more comfortable with the open quantifier response options, strongly preferring those response options to the options using number of visits. Respondents reported that the open quantifier response options were easier, and did not require remembering a concrete number of visits.

Similar to the second topic, for the third topic (related to an individual's experience with discussing cessation medications) and fourth topic (related to an individual's experience with discussing cessation methods other than medication), respondents were asked to answer two questions: one that used response options about the number of visits and one that used open quantifier response options. As with the second topic, respondents preferred the open quantifier response options. Respondents found it helpful that the questions included examples of cessation medications and cessation methods other than medication.

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

2016 update: interpretation of results from assessment of construct validity: 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. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample 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 0.0001 to evaluate the test results. P-values less than this this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results confirmed the hypotheses that the tobacco cessation rates are positively correlated with each other, suggesting they represent the same underlying quality construct of tobacco cessation care.

Interpretation of results from cognitive testing of survey questions: Respondents were able to answer the survey questions. They were able to interpret the addition of "tobacco" and they preferred open quantifier response options to response options that required them to recall the number of visits at which they discussed smoking or tobacco use cessation.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.1.1 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.

2b4.2. 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.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

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

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

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

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

2016 update: 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 .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However, the method can be used for comparison of any two measured entities.

Previous submission: Comparison of means and percentiles; analysis of variance against established benchmarks: if sample size is >400, we would use an analysis of variance.

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

2016 update: The p-value for commercial and Medicaid plans was <0.001 for all three rates, except for the commercial and Medicare Advising Smokers to Quit rates (p-value <0.01).

Table 3. HEDIS 2016 variation in performance across health plans

Product Line	Rate	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Commercial	Advising Smokers to Quit	132	75%	7%	66%	70%	75%	79%	83%	9%	0.009
	Discussing Cessation Medications	132	48%	8%	41%	43%	48%	52%	61%	9%	<0.001
	Discussing Cessation Strategies	131	44%	9%	34%	38%	42%	50%	58%	12%	<0.001
Medicaid	Advising Smokers to Quit	257	76%	6%	68%	73%	77%	79%	82%	6%	<0.001
	Discussing Cessation Medications	256	48%	8%	37%	43%	48%	54%	58%	11%	<0.001
	Discussing Cessation Strategies	256	43%	7%	34%	39%	44%	48%	52%	9%	<0.001
Medicare	Advising Smokers to Quit	55	86%	6%	78%	82%	86%	90%	93%	8%	0.006

EP: eligible population, the average denominator size across plans submitting to HEDIS
SD: standard deviation
IQR: interquartile range
p-value: p-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

Previous submission:
Commercial
ASTQ Rolling Average Rate
Data Element; 2010, 2009; 2008
N; 234; 14; 114
MEAN; 74.9; 79.5; 76.7
STDEV; 6.88; 5.99; 5.53

P10; 66.1; 73.8; 69.2
P25; 70.5; 73.9; 73
P50; 74.6; 79.8; 76.9
P75; 80.0; 84; 80.3
P90; 83.7; 87.7; 83.5

DSCM Rolling Average Rate
Data Element; 2010; 2009; 2008
N ; 231; 14; 115
MEAN; 50.5; 53.3; 54.4
STDEV; 7.8; 6.57; 7.29

P10; 40.3; 46.8; 45.2
P25; 45.3; 47.8; 48.7
P50; 50.3; 51.6; 53.6
P75; 55.5; 57.5; 61
P90; 62.0; 64.4; 63.6

DSCS Rolling Average Rate
Data Element; 2010; 2009; 2008
N; 228; 14; 115
MEAN; 42.8; 50; 49.7
STDEV; 8.7; 8.45; 7.46
P10; 33.0; 38.6; 39
P25; 35.9; 45.4; 44.6
P50; 41.2; 51; 49.7
P75; 47.8; 56.1; 55
P90; 55.1; 61.2; 59.8

Medicare
ASTQ Rolling Average Rate
Data Element; 2009
N; 295
MEAN; 78
STDEV; 7.91
STDERR; 0.46

MIN; 50
MAX; 100
P10; 67.7
P25; 73.3
P50; 78
P75; 82.9
P90; 87.8

Medicaid
ASTQ Rolling Average Rate
Data Element; 2010; 2009; 2008
N; 119; 99; 101
MEAN; 73.7; 74.3; 69.3
STDEV; 6.08; 5.3; 0.62
P10; 64.7; 67.1; 61.4
P25; 69.9; 70.8; 66.5
P50; 74.8; 74.9; 70.4
P75; 78.0; 77.7; 73.5
P90; 80.8; 80.8; 76.2

DSCM Rolling Average Rate
Data Element; 2010; 2009; 2008
N; 119; 99; 101
MEAN; 42.8; 43.4; 40.6
STDEV; 9.1; 9.38; 8.48

P10; 30.2; 29.4; 31.8
P25; 36.4; 37.2; 34.6
P50; 42.8; 43.4; 39.9
P75; 48.9; 51; 46.1
P90; 55.0; 56.6; 52.3

DSCS Rolling Average Rate
Data Element; 2010; 2009; 2008
N; 119; 98; 101
MEAN; 38.6; 38.8; 40.8
STDEV; 7.3; 7.78; 6.99

P10; 33.0; 28.4; 32.1
P25; 33.7; 34; 36.3
P50; 38.3; 38.3; 39.8
P75; 43.8; 44.4; 45.8
P90; 48.5; 50; 50.3

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

2016 update: The results above indicate that there is a six to 12 percent gap in performance between the 25th and 75th performing plans. For all product lines the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the commercial plans on the Discussing Cessation Strategies rate, which show a 12 percentage point gap between 25th and 75th percentile plans. The next largest gap is for Medicaid plans on the Discussing Cessation Medications, which show an 11 percentage point gap between 25th and 75th percentile plans. Additionally, the difference in performance between plans in the 10th and 90th percentiles varies from 12 to 27 points across the three rates and product lines. Overall, these results demonstrate that there are meaningful differences in performance for all three rates across commercial, Medicaid and Medicare product lines.

Previous submission: This information was not provided in the previous submission.

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

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

considered and pros and cons of each)
NA

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

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.**
Other
If other: Patient Survey

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**.
Patient/family reported information (may be electronic or paper)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
The data for this measure comes from a patient-reported survey.

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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
NCQA uses several mechanisms to solicit feedback from plans that participate in HEDIS reporting, including a Policy Clarification Support System and a HEDIS Users' Group. The Policy Clarification Support System allows NCQA to collect "real-time" feedback from measure users; through this system, NCQA receives thousands of inquiries each year on over 100 measures. The HEDIS Users' Group has 195 members for 2017; participation includes four conferences presented by NCQA to address key HEDIS implementation issues. NCQA has not heard about difficulties implementing this 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 Public Health/Disease Surveillance	Regulatory and Accreditation Programs HEDIS Health Plan Accreditation http://www.ncqa.org/programs/accreditation/health-plan-hp Quality Improvement (external benchmarking to organizations) Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx Quality Compass http://www.ncqa.org/tabid/177/Default.aspx

4a.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

ANNUAL 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 2015 the report included data from 814 HMOs and 353 PPOs, representing more than 171 million patients.

CMS HEALTH INSURANCE MARKET QUALITY RATING SYSTEM: This measure is used in the CMS developed, Quality Reporting Rating System (QRS) set of measures. The QRS measure set consists of measures that address areas of clinical quality management; enrollee experience; and plan efficiency, affordability and management. The measure set includes a subset of NCQA's HEDIS measures and one PQA measure.

HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings for Medicaid and Medicare health plan, which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS

measures among other factors. In 2012, a total of 455 Medicare Advantage health plans and 136 Medicaid health plans across 50 states were included in the rankings.

HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of commercial, Medicare Advantage and Medicaid health plans. In 2012, a total of 336 commercial health plans covering 87 million lives, 170 Medicare Advantage health plans covering 7.1 million Medicare beneficiaries, and 77 Medicaid health plans covering 9.1 million lives were accredited using this measure among others. Health plans are scored based on performance compared to benchmarks.

MEDICAID ADULT CORE SET: These are a core set of health quality measures for Medicaid-enrolled adults. The Medicaid Adult Core Set was identified by the Centers for Medicare & Medicaid (CMS) in partnership with the Agency for HealthCare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that adults enrolled in Medicaid receive nationally. Beginning in January 2014 and every three years thereafter, the Secretary is required to report to Congress on the quality of care received by adults enrolled in Medicaid. Additionally, beginning in September 2014, state data on the adult quality measures will become part of the Secretary's annual report on the quality of care for adults enrolled in Medicaid.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, 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.

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

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

4b. **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.
Between 2014 to 2016, the trend across all three rates and product lines has been stable or improved performance. For the Medicare product line, the average performance has improved two percentage points for the Advising Smokers to Quit rate. Average performance has remained stable on this rate for the commercial and Medicaid product lines. For the Discussing Cessation Medications rate, the average performance for the commercial product line has dropped one percentage point, and for the Medicaid product line it has improved one percentage point. For the Discussing Cessation Strategies rate, the average performance for both the commercial and Medicaid product lines has improved one performance point.

4c. Unintended Consequences

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

4c.1. **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.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
There were no identified unexpected benefits for this measure during testing or since implementation.

4d1.1. **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.

4d1.2. **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, as described in Section 3c1.

4d2.1. **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.

4d2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have not reported significant barriers to implementing this measure.

4d2.3. **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 Medicaid Adult Core Set and the CMS Health Insurance Market Quality Rating System

4d.3. **Describe how the feedback described in 4d.2 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.

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)**

0028 : Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention
2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
2803 : Tobacco Use and Help with Quitting Among Adolescents

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?

No

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

Refer to 5b.1

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.)

Answer for 5a.2: Identify differences, rationale, and impact on interpretability and data collection burden:

Preventive Care & Screening: Tobacco Use: Screening & Cessation Intervention (NQF #0028, stewarded by the AMA-convened Physician Consortium for Performance Improvement [AMA-PCPI]) evaluates the "percentage of patients 18 years and older who were screened for tobacco use one or more times within 24 months AND who received cessation counseling intervention if identified as tobacco user"; "cessation counseling intervention includes brief counseling (3 minutes or less), and/or pharmacotherapy." The denominator includes "patients aged 18 years and older seen for at least two visits or at least one preventive visit during the measurement period"; patients are excluded from the measure if there is "documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy)." It differs from NCQA's measure under review because it: 1) requires screening and intervention once every two years (rather than once every two); 2) includes only those patients who have had at least two visits or at least one preventive visit during the measurement period; and 3) excludes patients with a medical reason for not screening for tobacco use. Regarding the timing of screening and interventions, the USPSTF recommendation does not provide guidance about the frequency of screening or providing tobacco cessation interventions. Because of the harm caused by tobacco use and the positive outcomes associated with tobacco use cessation, NCQA has decided

to assess smoking and tobacco use cessation on an annual basis, rather than biannual basis. Regarding the visit requirement in the encounter, the AMA-PCPI measure is specified for individual clinicians and groups/practices, whereas the NCQA measure is specified for health plans. Because health plans may engage individuals in tobacco cessation outside of clinical visits, we chose not to require visits in the denominator. Lastly, regarding the medical reason exclusion, NCQA does not expect this type of exclusion to have a significant impact at the health plan level; therefore, we do not include this type of exclusion in the NCQA measure.

Tobacco Use and Help with Quitting Among Adolescents (NQF #2803, stewarded by the National Committee for Quality Assurance) evaluates the "percentage of adolescents 12 to 20 years of age during the measurement year for whom tobacco use status was documented and received help with quitting if identified as a tobacco user." There are no exclusions for the measure. It differs from NCQA's measure under review because it: 1) includes an evaluation of whether or not adolescents received tobacco screening; and 2) it focuses on adolescents rather than adults. It is specified for the clinician: group/practice level and EHR only. NCQA's measure under review focuses on evaluating whether patients who are current smokers or tobacco users receive information from their doctor or health provider about recommended cessation interventions. It also reports separate rates for the different recommended cessation interventions (advice, cessation medications, and cessation strategies), whereas the Tobacco Use and Help with Quitting Among Adolescents measure evaluates whether adolescents received any of the following: advice to quit, counseling on the benefits of quitting, assistance with or referral to a cessation support program, or current enrollment in a cessation program.

Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence (NQF #2600, stewarded by the National Committee on Quality Assurance) evaluates the "percentage of patients 18 years and older with a serious mental illness or alcohol or other drug dependence who received a screening for tobacco use and follow-up for those identified as a current tobacco user." There are two rates; rate 1 focuses on patients with a diagnosis of serious mental illness; rate 2 focuses on patients with a diagnosis of alcohol or other drug dependence. This measure is adapted from Preventive Care & Screening: Tobacco Use: Screening & Cessation Intervention (NQF #0028). There are no exclusions. It is specified at the health-plan level. The differences between NCQA's measure under review and this measure are similar as the differences between NCQA's measure under review and the AMA-PCPI measure, although this measure does not have exclusions. This measure is specified at the health plan measure, as is NCQA's measure under review; however, it focuses on a specific, at-risk population (patients with a serious mental illness or alcohol or other drug dependence).

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.

Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Medical Association & American Association for Physician Leadership
Andrew Baskin, MD, Aetna
Patrick Conway, MD, MSC, Center for Medicare & Medicaid Services
Jonathan D. Darer, MD, MPH, Medicalis
Helen Darling, Interim – National Quality Forum
Rebekah Gee, MD, MPH, FACOG, LSU School of Medicine and Public Health
Foster Gesten, MD, NY State Department of Health
David Grossman, MD, MPH, Group Health Physician
Christine S. Hunter, MD (Co-Chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare & Medicaid Services
Nancy Lane, PhD, Vanderbilt University Medical Center
Bernadette Loftus, MD, The Permanente Medical Group
Amanda Parsons, MD, Montefiore Health System
J. Brent Pawlecki, MD, MMM, The Goodyear Tire & Rubber Company
Susan Reinhard, PhD, RN, AARP Public Policy Institute
Eric C Schneider, MD, MSc, FACP (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH, Blue Shield of California
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

Smoking Cessation measure Workgroup: Steven Bernstein, Jonathan Foulds, Eric Heiligenstein, Corinne Husten, Carlos Jaen, Nancy Rigotti, Lowell Dale, Steve Schroeder, and David Warner

Cardiovascular MAP
Stephen D. Persell MD, MPH (Chair)
Kathy Berra, MSN, ANP, FAAN
David C. Goff, Jr., MD, PhD
Clarion Johnson, MD
Tom Kottke, MD
Eduardo Ortiz MD, MPH
Michael Pignone, MD, MPH
Randall S. Stafford, MD, PhD
Tracy Wolff, MD
Corinne Husten, MD, MPH
Samantha Tierney, MPH (Liaison)
Jason M. Spangler, MD, MPH, FACPM (Liaison)

The NCQA Smoking Cessation Measure Workgroup advised NCQA during measure development. They evaluated the way staff specified measures, assessed the content validity of measures, and reviewed field test results. As you can see from the list, the MAP consisted of a balanced group of experts, including representatives from primary care. Note that, in addition to the MAP, we also vetted these measures with a host of other stakeholders, as is our process. Thus, our measures are the result of consensus from a broad and diverse group of stakeholders, in addition to the MAP. The CVMAP advised on the first year analysis results.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 1997

Ad.3 **Month and Year of most recent revision:** 2009

Ad.4 **What is your frequency for review/update of this measure?** Approximately every 3-5 years or dependent on new guidelines or evidence.

Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:** For the survey instrument:
The CAHPS (R) program is funded and administered by the U.S. Agency for Healthcare Research and Quality, which works closely with a consortium of public and private organizations.

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.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0108
Corresponding Measures:
Measure Title: Follow-Up Care for Children Prescribed ADHD Medication (ADD)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: Percentage of children newly prescribed attention-deficit/hyperactivity disorder (ADHD) medication who had at least three follow-up care visits within a 10-month period, one of which is within 30 days of when the first ADHD medication was dispensed.
An Initiation Phase Rate and Continuation and Maintenance Phase Rate are reported.
Developer Rationale: Attention-deficit/hyperactivity disorder (ADHD) is a brain disorder marked by an ongoing pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development. Medications can improve function, but proper monitoring is recommended. The intent of this measure is to ensure timely and continuous follow-up visits for children who are newly prescribed ADHD medication. The goal is to encourage monitoring of children for medication effectiveness, occurrence of side effects and adherence.

Numerator Statement: Among children newly prescribed ADHD medication, those who had timely and continuous follow-up visits.
Denominator Statement: Children 6-12 years of age newly prescribed ADHD medication.
Denominator Exclusions: Children who had an acute inpatient encounter for mental health or chemical dependency following the Index Prescription Start Date
Children with a diagnosis of narcolepsy: Many of the medications used to identify patients for the denominator of this measure are also used to treat narcolepsy. Children with narcolepsy who are pulled into the denominator are then removed by the narcolepsy exclusion.
Children using hospice services during the measurement year. Children in hospice may not be able to receive the necessary follow-up care.

Measure Type: Process
Data Source: Claims (Only), Pharmacy
Level of Analysis: Health Plan, Integrated Delivery System

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **Most Recent Endorsement Date:** Mar 06, 2015

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 *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.

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 developer provided a rationale for timely follow-up visits after newly prescribing ADHD medication.
- The developer cited systematic reviews in the form of AAP clinical practice guidelines (**Strong recommendation; grade B** evidence) and AACAP practice parameters (**"minimal standard" recommendation; evidence graded rct and ut)** for the treatment of ADHD in children and adolescents.
- In the last review, the committee questioned the evidence supporting the 30-day timeframe and its linkage to improved outcomes, and noted barriers to meeting this requirement; the developer said the clinical guidelines support the 30-day period. The committee agreed that the measure addresses a high priority, as ADHD is one of the most prevalent behavioral health diseases in children.

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 states: "Numerous (>100) studies related to the care for patients with ADHD have been published since the publication of this guideline, none of which contradict the need for appropriate follow-up once treatment with medication begins."

Questions for the Committee:

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

Guidance from the Evidence Algorithm

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

The highest possible rating is 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.

- Performance data are summarized at the health plan level (commercial and Medicaid) for 2014-2016 for each of the 2 rates (initiation phase, continuance and maintenance [C&M] phase) reported in this measure. Data are summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. The developer provides denominator descriptions which include the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Performance scores

Plan	Year	Mean	Standard Deviation	10th Quartile	90th Quartile	Mean Eligible Population (per plan)
Initiation Phase						
Commercial	**2014**	39.1%	7.9%	29.1%	49.5%	447.0
	2015	37.5%	8.1%	27.3%	46.5%	438.9
	2016	39.0%	8.6%	29.1%	50.2%	396.9
Medicaid	**2014**	39.6%	11.3%	21.8%	53.0%	1,121.3
	2015	40.1%	10.8%	25.6%	54.0%	1,155.0
	2016	42.2%	11.0%	28.8%	55.5%	1,160.2
Continuance and Maintenance Phase						
Commercial	**2014**	45.9%	9.2%	35.1%	57.8%	185
	2015	44.7%	8.8%	35.1%	56.0%	177
	2016	46.8%	9.3%	35.6%	57.3%	163
Medicaid	**2014**	46.4%	15.1%	23.1%	63.1%	337
	2015	47.5%	15.6%	24.4%	65.2%	319
	2016	50.9%	13.3%	34.0%	67.3%	320

Disparities

- HEDIS measures are stratified by type of insurance. The developer cites literature that suggests that children from minority families experience decreased access to and utilization of health services for ADHD, as well as decreased rates of diagnosis of and treatment for ADHD.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is discussion needed regarding the relatively minimal change in these rates?*

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)

1.a. Evidence to Support Measure Focus

Comments:

**evidence is based on guidelines and practice parameters. but i question the evidence 'one month' association of improving outcome or decreasing harm (arbitrary; why not 1-2 months?). especially with the non-stimulants.

**I couldn't tell if the developer did an updated literature review. I found some studies that were relevant to understanding the measure but they did not reduce the evidence base. One article did suggest that the continuity measure reliability changes significantly depending on the follow-up rate and that the follow-up rate influences who gets into the measures

**Overview
First and most widely used child mental health quality measure. Measure was written and field tested from 2004-2005 (>12 years ago).
In the past, there was a sentiment that there should be support for this measure because it was one of the few available that was specific to children with mental health problems. However, perhaps it's time to re-examine whether HEDIS measures, developed originally for accreditation purposes, are suitable for use as national quality measures? Does the approach NCQA use for measure development, and testing reliability and validity sufficient to support use for public reporting on quality in publicly funded care?

The data source is administrative data from health plans that report on this HEDIS measure.
Process measure With the exception of atomoxetine, ADHD medication includes stimulants, a Schedule II medication requiring a triplicate rx.

Evidence to support (1a)
Based on AAP Clinical Practice Guideline and AACAP practice parameter. Both guidelines however comment on how the frequency of the follow-up visit depends on the clinical characteristics.

AAP: "subsequent visits "WILL DEPEND ON THE RESPONSE", and the minimum of at least 2 visits/year is contingent on the child's clinical status: "UNTIL IT IS CLEAR THAT TARGET GOALS ARE PROGRESSING AND STABLE"

AACAP: Note: this 2007 practice parameter is not current, and is listed as historical on AACAP practice parameter website. Nevertheless, it too emphasizes clinical characteristics of the child should influence frequency of stimulant medication safety and monitoring visits. "The frequency and duration of follow-up sessions should be individualized…depending on the severity of ADHD symptoms, the degree of comorbidity, response to treatment, degree of impairment."
NCQA (p5, evidence 1a) acknowledges that the evidence supporting this measure is the benefits of consistent medication treatment and "timely" follow-up.

However:
1)Is a minimum of 2 follow-up visits during a 9 month time period of which one can be a telephone call acceptable medication safety monitoring for children ages 6-12 years on a stimulant medication, especially with relatively high pharmacy fill persistence? Note: in the AACAP practice parameters the method of follow-up is not specified but the assumption was face to face visit with the child psychiatrist. Note: p6 (evidence 1a), NCQA reports "Significant contact with the clinician should typically occur 2-4 times per year in cases of uncomplicated ADHD and up to weekly sessions at times of severe dysfunction or complications of treatment." However, I could not find support for this statement in the 2007 AACAP parameter.

2)Why is a telephone visit acceptable for medication monitoring visit if recommendations include monitoring BP, p and ht/weight?

3)If continuity of care is the "key to effective long term management of a patient with ADHD" (p6, evidence 1a), then why isn't the main care process to be measured continuity of stimulant medication treatment? It is assumed that very few physicians will provide a 30 day refill on stimulant medication if the child fails to come to medication follow-up visits during a 9 month "continuation and maintenance" time period. This is important because the average #follow-up visit for ADHD/6 months in publicly-funded primary care is one. Though on average higher in community mental health care (5), only 28% of children receiving ADHD care receive any stimulant medication treatment (Zima et al. 2010.)

4) Although, the adherence rates vary across percentiles, the main finding of little change over years on this measure and potential overestimation of adherence for C&M make this less than meaningful as a quality indicator that has stimulated much change in quality of care for children with ADHD who are treated with stimulant medication.

5) The main study supporting this measure is the MTA study from 1999, of which the majority of children were White from 2 parent, often both college educated, families receiving care from the study sites.

**Regarding Maintenance measure and new information: There has been a large shift in patient-provider interaction to follow more technology-based methods - eMR 'chats', apps, direct emails to physicians, etc. This measure limits initial follow up visit as face-to-face within 30 days of starting ADHD medication. Second change in patient care that is continuing to expand is the growth

of high deductible plans. This has a direct impact on patient's out-of-pocket spend that has increased physician's desire to find alternative methods of patient communication (many unwilling to be pay for another office copay within 30 days of first one).

**Process measure: appropriate follow-up leads to better outcomes.

1.b. Performance Gap

Comments:

** disparities noted especially with minority populations.

**Yes. There was significant differences cited between the 25th and 75 percentile in performance on the measure. The average performance was low. There was no data by subgroup presented other than Medicaid and commercial. There was some literature cited that indicates racial disparities in ADHD access and quality of treatment.

**Little meaningful change in adherence rates by insurance type over many years.

**Concern on why no change in the initiation phase f/u for commercial and really negligible change in continuation phase - any work being done to determine if this is truly a sign of poor quality or alternate patient engagement services?

**Much variability and disparities exist; there is great opportunity for improvement.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Claims (Only), Pharmacy

Specifications:

- The measure is specified for the health plan level of analysis and stratified by product line (commercial or Medicaid).
- The measure is based on administrative claims.
- A higher score indicates better quality.
- The measure includes 2 rates:
  - Initiation phase rate (follow-up within 30 days of new prescription)
  - Continuation and maintenance (C&M) phase rate (Compliant for Initiation Rate AND have at least 2 more follow-up visits from 31-300 days after initial prescription.)
- The numerator for the initiation rate is defined as an outpatient, intensive outpatient, or partial hospitalization follow-up fixit with a practitioner with prescribing authority, within 30 days after the earliest prescription dispensing date for a new ADHD medication. The numerator is identified by certain code combinations.
  - Developer has provided value sets for codes separately.
- The numerator for C&M rate is children who are compliant for Rate 1 AND have documentation of at least two follow-up visits with any practitioner from 31–300 days (9 months) after the earliest prescription dispensing date for a new ADHD medication. One of the two visits may be a telephone visit. Follow-up visits may be identified using certain code combinations.
- The developer provides instructions for identification of the initiation rate denominator and the C&M rate denominator.
- Denominator exclusions include:

- Children with an acute inpatient encounter for mental health or chemical dependency following the index prescription start date,
- Children with narcolepsy
- Children using hospice during the measurement year.

- A calculation algorithm is provided.
- There have been no changes to the specification since the last endorsement.

Questions for the Committee:

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

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Previous measure score reliability testing was calculated from HEDIS data using a signal-to-noise analysis.

Describe any updates to testing:

- The developer provided a 2016 update of measure score reliability testing using a signal-to-noise analysis.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Testing included use of a signal-to-noise analysis.
- The developer provides data on the testing sample, including the number of health plans and the mean or median eligible population per plan.

Results of reliability testing

Beta-Binomial Statistic For Each Measure Rate: Mean Reliability

	Commercial		Medicaid	
Year	**2013**	**2016**	**2013**	**2016**
Initiation Phase	0.71	0.90	0.93	0.98
C&M Phase	0.66	0.75	0.92	0.95

- The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Generally, a reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities.

Questions for the Committee:

- o *Does the Committee think there is a need to re-discuss and re-vote on reliability?*
- o *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Specifications are precise (Box 2)→empirical reliability testing (Box 4)→score level testing (Box 5)→signal-to-noise analysis shows high signal strength (Box 6)→High

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence: The evidence does not specify an optimal frequency for follow-up visit timeframes.

Question for the Committee:

- o *Do you agree with the timeframe for follow-up visits and the number of such visits?*

2b2. Validity testing

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

For maintenance measures, summarize the validity testing from the prior review:

- The developer previously provided both construct validity testing and face validity. In 2013, construct validity was calculated. Face validity was assessed in 2004, with four panels of experts from diverse backgrounds.
- In the previous phase, Committee members questioned the evidence supporting the 30-day timeframe and its linkage to improved outcomes. The developer noted the timeframe is supported by AAP and AACAP guidelines and that they considered other timeframes ranging from 15-45 days worked best in terms of access and claims processing issues.
- The Committee also expressed concern that the exclusion of those who are non-compliant with the 30-day follow-up are the individuals who might need follow-up care the most. The developer stated this measure addresses the aspect of ADHD care related to follow-up and focuses on monitoring response to medication.

Describe any updates to validity testing:

- There are no updates to validity testing.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity**
- ☒ **Empirical validity testing of the measure score**

Validity testing method:

- In 2014, the developers tested for construct validity to determine whether the rates of this measure were correlated with another HEDIS measure, Children and Adolescents Access to Primary Care Practitioners. They hypothesized the two measures would be positively correlated, as organizations that perform well on providing follow-up care for children on an ADHD medication should also perform well on providing access to primary care practitioners. They used a Pearson correlation test to estimate the strength of the association.
- Face validity was assessed with four panels of experts. In this process, NCQA uses a standardized process called the HEDIS measure life cycle, which included field testing.

Validity testing results:

For face validity, the measure was deemed to have the desirable attributes of a HEDIS measure in 2005 (including relevance, scientific soundness, and feasibility).

For construct validity, the developers calculated Pearson correlations as follows:

Pearson Correlation Coefficients between *Follow-up Care for Children Prescribed ADHD Medication* and *Child and Adolescent Access to Primary Care Practitioners*: Commercial Plans, 2013

	Access to Primary Care Practitioners
ADHD Follow-up: Initiation Rate	0.4
ADHD Follow-up: Continuation & Maintenance Rate	0.4

All correlations are significant at p<.05

Pearson correlations measure the degree of association between two quantitative variables. For the social sciences, scores of 0.37 or larger are considered to have a "large" correlation effect. (Medium effect is 0.24 – 0.36 and small effect is 0.10 – 0.23.)

Questions for the Committee:

- o *How does the evidence support the number of follow-up visits?*
- o *No updated testing information is presented. The prior testing demonstrated good validity. Does the Committee think there is a need to re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Exclusions include:
  - o Children with an acute inpatient encounter for mental health or chemical dependency following the index prescription start date
  - o Children with narcolepsy
  - o Children using hospice during the measurement year.
- No data are provided on frequency or testing of exclusions.
- The developer notes that ICD-9/10 CM Diagnosis codes for narcolepsy are used for excluding patients diagnosed with narcolepsy, stating that "[w]hile these codes have not been tested in the context of this measure for validity, they are widely used across practitioners and considered to be valid."

- The developer states that this measure does not allow for exclusions for patient refusal, provider refusal, or un-specified exclusions.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Have the threats to validity related to exclusions been adequately addressed?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

NCQA calculates an inter-quartile range (IQR) for each indicator, which 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 which calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The statistic is then compared against a normal distribution. If the p-value of the test statistic is less than .05, then the two plans' performances are significantly different from each other. Using this method, NCQA compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. These are updated from 2013 performance rates. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The p-value for all product lines and rates was <.001 except form the Commercial health plan rate for C&M rate (p-value 0.002).

ABILITY TO IDENTIFY STATISTICALLY SIGNIFICANT/MEANINGFUL DIFFERENCES

HEDIS 2016 Variation in Performance across Health Plans

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

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

Question for the Committee:

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

2b6. Comparability of data sources/methods:

Not needed

2b7. Missing Data

- The developer states that plans collect this measure using all administrative data sources, and asserts that NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

Guidance from the Validity Algorithm

Specifications are consistent with evidence (Box 1)→potential threats mostly assessed (no data about exclusions) (Box 2) →empirical reliability testing (Box 3)→score level testing (Box 6)→correlation effect high (0.4 Pearson) (Box 7) →High certainty (Box 8a) →High

The highest possible rating is HIGH.

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

Committee pre-evaluation comments

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

2a.1 & 2b.1 Specifications: Reliability Specifications

Comments:

**no additional concerns

**There seems to be missing codes. It's not clear that the new update to the psychiatric CPT codes are reflected in the value set such as 90862 (pharmacological management). It not clear whether or how telephonic follow-up is captured.

**The NCQA definition of validity testing includes that "the measure score correctly reflects the quality of care provided". (For outcome measures, includes clinically meaningful differences in performance).

The data source is administrative data from health plans that participate in NCQA accreditation, and thus report on adherence to HEDIS measures. Exclusion and inclusion criteria are appropriate, and measure is well specified. The approach to counting days supplied for overlapping rx's is reasonable.

**Similar concern as above, regarding consistent implementation due to patient portals allowing eMR 'chats' with physicians, and other alternate ways to engage patients; especially for parents of children not experiencing any immediate side effects, weekly titration plan went well at home with follow up calls to physician's office as needed, etc.

**Data elements clearly defined.

2a.2 Reliability Testing

Comments:

**no additional concerns.

**Reliability testing consisted of signal-to-noise ratios. It appears to have been conducted only on the score level. The signal to noise ratio was considered good. No other reliability testing was conducted.

**There is no data on reliability or clinical validity at the agency, program or provider level.

The sample of health plans for Initiation phase was 344 commercial and 180 Medicaid. For the C & M phase, 61% of the commercial health plans (210/344) corresponding to 41% eligible cases (163/397), and 83% of the Medicaid health plans (151/180) corresponding to only 28% eligible cases (320/1160).

Reliability was solely based on beta-binomial method given limited to health plan level data, and the proportion of total variation that could be attributable to a health plans was at least adequate (>-.7; range: I: .75-.90; C&M: ,98-,95)

**good.

**Reliability testing- measure score- conducted with the data source. Reliability is high.

2b.1 Validity Specifications

Comments:

**no additional concerns

**The validity is largely based on face validity, construct validity and clinical practice guidelines. I don't have validity concerns.

**see above concerns re: whether a telephone visit is acceptable to monitor stimulant medication safety in children as well as the bare minimum # of "visits"/9 month C& M phase despite national recommendations that # be aligned with child's clinical need.

It does not directly report general continuity of care for ADHD medication treatment.

**Target population relies on parents to take them to doctor's office. Limiting initiation phase to face-to-face is not optimal for many parents.

**The evidence doesn't clearly specify what the optimal frequency is for follow-up visit time frames.

2b.2 Validity Testing

Comments:

**no additional concerns

**Validity was tested through face validity and construct validity (correlation between the measure and a measures of children and adolescent access to primary care measure).

**Validity testing based on NCQA approach to assess face validity as part of their HEDIS measure life cycle. This approach also includes assessment of testing construct validity using adherence from a HEDIS measure that assesses any contact with a PCP. From 2013 validity testing, the pearson correlation coefficients if 0.4 (moderate, .3-.5)) for commercial health plan and <=.3 for Medicaid health plans which would correspond to low correlation (.1-.3) Nevertheless, NCQA emphasizes significance testing, but I could not locate the p values in section 2b.2.4 where the statement is, "These results indicate that the follow0up care for children with a rx for ADHD medication measure is a valid measure of a plan's quality" (p13, 2b.2.4).

There is no data supporting the clinical validity of this measure. What is the outcome for children that don't meet vs. don't meet the C&M phase criteria for continuous medication treatment? No additional validity testing has been done. Instead the same approach is used for 2016 health plan data as was done with 2013 health plan data.

There is no capacity for risk adjustment, even though both clinical guideline/practice parameter support frequency of medication follow-up visit on child's response to treatment and clinical severity.

"Meaningful differences" in performance is based on statistically significant differences in rates of adherence to measure between the 25th percentile and 75th percentile.

However, overall the average adherence rates in 2016 for commercial plans are I: 39%, C&M: 46.8%; and for Medicaid: I: 42.2%; C&M: 50.9% with very little change when compare to average rates in 2013 (commercial: I: 38.4%, C&M: 45.3%; Medicaid: I: 39.1%; C&M: 45.2)

Note: the C&M adherence rate is overestimated because it excludes a large proportion of children who do not meet the criteria for "continuous medication treatment" (total 90 day gap in rx filled/10 months).

Although, the adherence rates vary across percentiles, the main finding of little change over years on this measure and potential overestimation of adherence for C&M make this less than meaningful as a quality indicator that has stimulated much change in quality of care for children with ADHD who are treated with stimulant medication.

Note: Measure Worksheet, p4 follow-up "fixit" instead of visit; p6 states that a Pearson r >.37 is "large" but found different definition from other sources.

Note: On the original behavioral health measurement advisory panel (p55, measure worksheet), it appears that only one "expert" is a child and adolescent psychiatrist. Of the MD's, Dr. John Strauss(founder, Beacon Health Options) is assumed to be a CAP. Katherine Bradley is an internist and Dr. Korsen is a family medicine specializing in geriatrics per internet search.

NCQA reports that "no modifications to this measure have been required based on feedback received." (p9, measure worksheet). However, their approach to gathering information from public comments is often limited to 3 responses: agree, disagree or support with modifications with limited character space for discussion.

**ok.

**Face and empirical validity testing of the measure score.

2b.3.-2b7. Testing (Related to Potential Threats to Validity)
Comments:
**no additional concerns

**I had trouble finding the developer instructions for addressing continuous enrollment. Continuous enrollment will affect the generalizability of the measure.

**The problem with relatively high attrition for the C&M phase appears minimized. The missing data problem and selection bias for the C&M phase is minimized: "NCQA's audit process checks that plans' measure calculations are not biased due to missing data." (p21, evidence 2a, 2b).

**ok.

**Exclusions are appropriate. Risk adjustment: none Validity-moderate to high

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 required data elements are routinely collected and are available in electronic sources.
- The developer did not report any implementation challenges.

Questions for the Committee:
- *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:
**electronic sources; no additional concerns.

**The main limitation is that the data source is health plan-reported data. The challenges of implementation for publicly-funded programs at the state, agency or provider level are not addressed.

**ok.

**Required data is routinely collected. Feasibility is moderate to high.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:

- This measure is used in the following CMS programs:
  - Medicaid Child Core Set
  - Electronic Health Record Incentive Program
  - The measure is used in the following; *NOTE: all are being phased out by 12/31/18 and replaced by MIPS.*
    - CMS Physician Quality Reporting System (PQRS)
    - Physician Feedback/Quality and Resource Use Reports (QRUR)
    - Physician Value-Based Payment Modifier (VBM)
- This measure is used by NCQA for scoring in accreditation:
  - In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries.
  - The measure is also used in the Accountable Care Organization Accreditation Program.
- This measure is also reported and used in the following:
  - State of Health Care Annual Report (nationally and by geographic region)
  - Quality Compass
  - Qualified Health Plan (QHP) Quality Rating System (QRS).
- This measure is used to calculate health plan rankings which are reported in Consumer Reports and on the NCQA website.

Improvement results:

- As shown in the performance results, the performance rates been generally stable across commercial and Medicaid plans, from 2014-2016.
- Across both commercial and Medicaid plans, there continues to be variation between the 10th and 90th percentiles.

Unexpected findings (positive or negative) during implementation: None identified during testing or implementation.

Potential harms: None identified during testing or implementation.

Vetting of the measure:

- The developer notes they use several methods to obtain input, including several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer notes that health plans that report HEDIS calculate their rates and so know their performance when submitting to NCQA, and that NCQA publicly reports rates across all plans so that the plans can understand their relative performance.
- While the developers provide technical assistance for calculating/implementing the measure, It is not clear whether the developers provide specific technical assistance with interpreting the results.

Feedback:

- In 2016, the MAP Medicaid Task Force again supported the measure's continued use in the Medicaid Child Core Set.
- The developer reported that health plans have not reported significant barriers to implementing the measure, and no modifications to this measure have been required based on feedback received.

***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

4. Usability and Use:

Comments:

**I would like to see additional evidence (other than guidelines and parameters) that 1 month face to face follow-up improves outcome or decreases harm. especially with telemedicine becoming more common, and if vitals can be obtained at other places) (and other variables unique to individual patient/family circumstances).

-I question why strattera is included, especially since it can take longer than 1 month to see effect.

-why just 6-12 yo? (why not all ages?)

**The measures is being widely implemented as part of a number of CMS programs and in HEDIS. The developer did not provide feedback on how providers were using the measure to improve care and if they, or health plans, found it useful.

** The measure is reported to be used in NCQA products as well as several CMS quality monitoring and reporting activities.

Note: On the original behavioral health measurement advisory panel (p55, measure worksheet), it appears that only one "expert" is a child and adolescent psychiatrist. Of the MD's, Dr. John Strauss(founder, Beacon Health Options) is assumed to be a CAP. Katherine Bradley is an internist and Dr. Korsen is a family medicine specializing in geriatrics per internet search.

NCQA reports that "no modifications to this measure have been required based on feedback received." (p9, measure worksheet). However, their approach to gathering information from public comments is often limited to 3 responses: agree, disagree or support with modifications with limited character space for discussion.

** The performance results have not changed much over past 3 years for commercial membership - should do an assessment of why no change. Make sure this truly is due to no follow up care is being provided versus other forms of follow up care, that is more patient-friendly, is being provided.
For physician groups, unintended consequence is being given a report card for many of its pay-for-performance contracts that show a poor score for this metric when in reality there was much follow up care performed, just not always face-to-face. Also, it would be interesting to understand if there is a higher capture rate if 4 to 6 weeks post-initial dose start improves metrics without losing any quality of care.

** Data is publicly available and current use is in an accountability program.

Criterion 5: Related and Competing Measures

Related or competing measures

- N/A

Harmonization

- N/A

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☒ **Yes** ☐ **No**

Pre-meeting public and member comments

- No comments received.

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

Measure Title: Follow-Up Care for Children Prescribed ADHD Medication

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: 12/2/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*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: Follow-Up Care for Children Prescribed ADHD Medication

☐ 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.12 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.

Children newly prescribed attention-deficit hyperactivity disorder (ADHD) medication >> timely follow-up visits occur >> medication effectiveness and any adverse effects are assessed >> dose is adjusted if needed >> treatment adherence and health outcomes are improved

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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

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

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

☑ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

Table 1: American Academy of Pediatrics Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• ADHD: Clinical Practice Guideline for the Diagnosis, Evaluation, and Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents • American Academy of Pediatrics (AAP) • 2011 • ADHD: Clinical Practice Guideline for the Diagnosis, Evaluation, and Treatment of Attention-Deficit/Hyperactivity Disorder in Children and Adolescents. Subcommittee On Attention-Deficit/Hyperactivity Disorder, Steering Committee On Quality Improvement And Management Pediatrics Oct 2011, peds.2011-2654; DOI: 10.1542/peds.2011-2654 • http://pediatrics.aappublications.org/content/early/2011/10/14/peds.2011-2654
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.	American Academy of Pediatrics Clinical Practice Guideline for the Diagnosis, Evaluation and Treatment of ADHD in Children and Adolescents *Action Statement 4*: The primary care clinician should recognize ADHD as a chronic condition and, therefore, consider children and adolescents with ADHD as children and youth with special health care needs. Management of children and youth with special health care needs should follow the principles of the chronic care model and the medical home. *Grade B: Strong Recommendation* Additionally, in the supplemental information provided along with the 2011 AAP Guideline, the following recommendations are made: • "A face-to-face follow-up visit is recommended by the fourth week of medication, during which clinicians review the responses to the varying doses and monitor adverse effects, pulse, blood pressure, and weight."

<table>
<tr><td></td><td>• “Subsequent visits will depend on the response but should occur at least 2 times per year, until it is clear that target goals are progressing and stable, and then periodically as determined by the family and the clinician.”</td></tr>
<tr><td>Grade assigned to the evidence associated with the recommendation with the definition of the grade</td><td>The grade assigned by the AAP to the evidence supporting the listed Clinical Practice Guidelines for the Diagnosis, Evaluation and Treatment of ADHD in Children and Adolescents was a grade of B.
Grade B: RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies</td></tr>
<tr><td>Provide all other grades and definitions from the evidence grading system</td><td>
<table>
<tr><th>Evidence Quality</th><th>Preponderance of Benefit or Harm</th><th>Balance of Benefit and Harm</th></tr>
<tr><td>A. Well-designed RCTs or diagnostic studies on relevant population</td><td>Strong recommendation</td><td rowspan="3">Option</td></tr>
<tr><td>B. RCTs or diagnostic studies with minor limitations; overwhelmingly consistent evidence from observational studies</td><td rowspan="2">Recommendation</td></tr>
<tr><td>C. Observational studies (case-control and cohort design)</td></tr>
<tr><td>D. Expert opinion, case reports, reasoning from first principles</td><td>Option</td><td>No Rec</td></tr>
<tr><td>X. Exceptional situations in which validating studies cannot be performed and there is a clear preponderance of benefit or harm</td><td>Strong recommendation
Recommendation</td><td></td></tr>
</table>
</td></tr>
<tr><td>Grade assigned to the recommendation with definition of the grade</td><td>The AAP assigned the categorization of Strong Recommendation to the listed Clinical Practice Guidelines for the Diagnosis, Evaluation and Treatment of ADHD in Children and Adolescents.
Strong Recommendation: A strong recommendation means that the committee believes that the benefits of the recommended approach clearly exceed the harms of that approach (or, in the case of a strong negative recommendation, that the harms clearly exceed the benefits) and that the quality of the evidence supporting this approach is either excellent or impossible to obtain. Clinicians should follow such guidance unless a clear and compelling rationale for acting in a contrary manner is present</td></tr>
<tr><td>Provide all other grades and definitions from the recommendation grading system</td><td>American Academy of Pediatrics Grading System
• Recommendation: A recommendation means that the committee believes that the benefits exceed the harms (or, in the case of a negative recommendation, that the harms exceed the benefits), but the quality of the evidence on which this recommendation is based is not as strong. Clinicians also generally should follow such guidance but also should be alert to new information and sensitive to patient preferences
• Option: An option means either that the evidence quality that exists is suspect or that well-designed, well-conducted studies have demonstrated little clear advantage to one approach versus another. Options offer clinicians flexibility in their decision-making regarding appropriate practice, although they may set boundaries on alternatives. Patient preference should have a substantial role in influencing clinical decision-making, particularly when policies are expressed as options.
• No Recommendation: No recommendation is made when there is both a lack of pertinent evidence and an unclear balance between benefits and harms. Clinicians should feel little constraint in their decision-making when addressing areas with insufficient evidence. Patient preference should have a substantial role in influencing clinical decision-making.</td></tr>
</table>

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Guidelines from the American Academy of Pediatrics (Wolraich et al. 2011) cite a randomized control trial of 600 children diagnosed with ADHD ages 7-9 years old, as well as a prospective observational cohort study of 34 children. It also cites two systematic literature reviews and the chronic care model (Bodenheimer, Wagner, & Grumbach 2002). The action statement also received a grade of B, which indicates high quality evidence including randomized controlled trials. The evidence supporting the AAP guidelines received a grade of B, indicating that the guideline is supported by strong evidence consisting of randomized controlled trials and that there is a preponderance of benefit compared to harm. Additionally, the AACAP guideline is based on 15 randomized controlled trials in addition to multiple controlled and uncontrolled trials, all providing evidence of the efficacy of continuous medication treatment and exploring side effects and other aspects of ADHD medication use that require monitoring.
Estimates of benefit and consistency across studies	The evidence supporting the guidelines demonstrate the benefits of consistent treatment, side-effect monitoring and medication adjustment for children and adolescents with ADHD. Timely follow-up visits ensure children and adolescents on AHDH medications receive these services. Both the AAP and AACAP guidelines are based on the Multi-Modal Treatment Study of Children with ADHD (MTA) (in addition to other studies). In the MTA study, children with ADHD were randomized to four groups: algorithmic medication treatment alone, psychosocial treatment alone, a combination of algorithmic medication management and psychosocial treatment, and community treatment. Algorithmic medication treatment consisted of monthly appointments in which the dose of medication was titrated according to parent and teacher rating scales. Children in all four treatment groups demonstrated benefits to treatment in terms of reduced symptoms of ADHD compared to baseline. The two groups that received algorithmic medication management showed a superior outcome with regard to ADHD symptoms compared with those that received intensive behavioral treatment alone or community treatment (MTA Cooperative Group, 1999a [rct]). Once the study treatments ceased at 14 months, the combined and medication groups lost some of their treatment gains, in part because of medication discontinuation and in part because the medication was now being given in the community with less careful monitoring and dose adjustment (MTA Cooperative Group, 2004a [rct], 2004b [rct]).
What harms were identified?	The American Academy of Pediatrics provided an analysis of net benefit and concluded that there is a preponderance of benefit over harm because of the opportunity to assess adverse effects of medication and to sustain treatment.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous (>100) studies related to the care for patients with ADHD have been published since the publication of this guideline, none of which contradict the need for appropriate follow-up once treatment with medication begins.

Table 2: American Academy of Child and Adolescent Pyschiatry Guidelines

Source of Systematic Review: • **Title** • **Author**	• Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. • American Academy of Child and Adolescent Psychiatry (AACAP) • 2007

• **Date** • **Citation, including page number** • **URL**	• Practice Parameter for the Assessment and Treatment of Children and Adolescents With Attention-Deficit/Hyperactivity Disorder • Pliszka, Steven. Journal of the American Academy of Child & Adolescent Psychiatry , Volume 46 , Issue 7 , 894 – 921. • http://www.jaacap.com/article/S0890-8567(09)62182-1/abstract
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 Guideline* The key to effective long-term management of the patient with ADHD is continuity of care with a clinician experienced in the treatment of ADHD. The frequency and duration of follow-up sessions should be individualized for each family and patient, depending on the severity of ADHD symptoms; the degree of comorbidity of other psychiatric illness; the response to treatment; and the degree of impairment in home, school, work, or peer-related activities. The clinician should establish an effective mechanism for receiving feedback from the family and other important informants in the patient's environment to be sure symptoms are well controlled and side effects are minimal. Although this parameter does not seek to set a formula for the method of follow-up, significant contact with the clinician should typically occur two to four times per year in cases of uncomplicated ADHD and up to weekly sessions at times of severe dysfunction or complications of treatment. *Specific Recommendations* *Recommendation 6*: A Well-Thought-Out and Comprehensive Treatment Plan Should Be Developed for the Patient With ADHD. The treatment plan should be reviewed regularly and modified if the patient's symptoms do not respond. *Minimal Standard [MS]* *Recommendation 9*. During a Psychopharmacological Intervention for ADHD, the Patient Should Be Monitored for Treatment-Emergent Side Effects. *Minimal Standard [MS]* *Recommendation 12*. Patients Should Be Assessed Periodically to Determine Whether There Is Continued Need for Treatment or If Symptoms Have Remitted. Treatment of ADHD Should Continue as Long as Symptoms Remain Present and Cause Impairment. *Minimal Standard [MS]*
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The grade assigned by AACAP to the evidence supporting the listed Practice Parameters for the Assessment and Treatment of Children and Adolescents with ADHD varied by study and included [rct] and [ut]: • [rct] Randomized, controlled trial is applied to studies in which subjects are randomly assigned to two or more treatment conditions. • [ut] Uncontrolled trial is applied to studies in which subjects are assigned to one treatment condition
Provide all other grades and definitions from the evidence grading system	The grades assigned by the AACAP to evidence supporting the Practice Parameters for the Assessment and Treatment of Children and Adolescents with ADHD noted the strength of the study by listing the study type. Other studies used to support the guidelines included: • [rct] Randomized, controlled trial is applied to studies in which subjects are randomly assigned to two or more treatment conditions. • [ct] Controlled trial is applied to studies in which subjects are nonrandomly assigned to two or more treatment conditions.

	• [ut] Uncontrolled trial is applied to studies in which subjects are assigned to one treatment condition. • [cs] Case series/report is applied to a case series or a case report.
Grade assigned to the **recommendation** with definition of the grade	AACAP assigned a grade of *[MS] Minimal Standard* to the listed Practice Parameters for the Assessment and Treatment of Children and Adolescents with ADHD. *[MS] Minimal Standard* is applied to recommendations that are based on rigorous empirical evidence (e.g., randomized, controlled trials) and/or overwhelming clinical consensus. Minimal standards apply more than 95% of the time (i.e., in almost all cases).
Provide all other grades and definitions from the recommendation grading system	*American Academy of Child and Adolescent Psychiatry Grading System* • *[CG] Clinical Guideline is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized, controlled trials) and/or strong clinical consensus. Clinical guidelines apply approximately 75% of the time (i.e., in most cases).* • *[OP] Option is applied to recommendations that are acceptable based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion, but lack strong empirical evidence and/or strong clinical consensus.* • *[NE] Not Endorsed is applied to practices that are known to be ineffective or contraindicated.*
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Guidelines from the American Academy of Child and Adolescent Psychiatry (Pliszka 2007) cite two randomized control trials, one of which enrolled 600 children diagnosed with ADHD ages 7-9 years. The other study enrolled 103 children diagnosed with ADHD also ages 7-9. In addition to these large randomized control trials, each recommendation cited additional studies which provided further evidence examining treatment planning for children with ADHD, side effects associated with ADHD medications, medication adherence, and treatment adjustment. • Recommendation 6: Treatment Plan review and modification: 3 RCTs • Recommendation 9: Side Effect Monitoring: 4 RCTS, 1 CT, 1 UT • Recommendation 12: Periodic Assessment of Symptoms: 8 RCTs, 5 UTs Overall, the quality of the evidence regarding follow-up care for children with a prescription for an ADHD medication is good. Guidelines from the American Academy of Pediatrics and the American Academy of Child and Adolescent Psychiatry cite the Multi-Modal Treatment Study of Children with ADHD (MTA). The 1999-published MTA study, sponsored by the National Institute of Mental Health, was a randomized control trial, multi-site study of nearly 600 elementary school children, 7-9 years of age who were diagnosed with ADHD and randomly assigned to one of four treatment modes: medication alone; psychosocial/behavioral treatment alone; a combination of both; or routine community care. The MTA study demonstrated that, on average, carefully monitored medication management with monthly follow-up is more effective than intensive behavioral treatment for ADHD symptoms, for periods lasting as long as 14 months. The quality of this study can be considered high due to the randomization and large sample size. The AACAP guideline is based on 15 randomized controlled trials in addition to multiple controlled and uncontrolled trials, all providing evidence of the efficacy of continuous

	medication treatment and exploring side effects and other aspects of ADHD medication use that require monitoring.
Estimates of benefit and consistency across studies	The evidence supporting the guidelines demonstrate the benefits of consistent treatment, side-effect monitoring and medication adjustment for children and adolescents with ADHD. Timely follow-up visits ensure children and adolescents on AHDH medications receive these services. Both the AAP and AACAP guidelines are based on the Multi-Modal Treatment Study of Children with ADHD (MTA) (in addition to other studies). In the MTA study, children with ADHD were randomized to four groups: algorithmic medication treatment alone, psychosocial treatment alone, a combination of algorithmic medication management and psychosocial treatment, and community treatment. Algorithmic medication treatment consisted of monthly appointments in which the dose of medication was titrated according to parent and teacher rating scales. Children in all four treatment groups demonstrated benefits to treatment in terms of reduced symptoms of ADHD compared to baseline. The two groups that received algorithmic medication management showed a superior outcome with regard to ADHD symptoms compared with those that received intensive behavioral treatment alone or community treatment (MTA Cooperative Group, 1999a [rct]). Once the study treatments ceased at 14 months, the combined and medication groups lost some of their treatment gains, in part because of medication discontinuation and in part because the medication was now being given in the community with less careful monitoring and dose adjustment (MTA Cooperative Group, 2004a [rct], 2004b [rct]). In terms of side-effect monitoring, the AACAP guidelines are based on four randomized controlled trials, one controlled trial and one uncontrolled trial. These trials found that it is prudent to monitor side effects in order to optimize patient outcomes.
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous (>100) studies related to the care for patients with ADHD have been published since the publication of this guideline, none of which contradict the need for appropriate follow-up once treatment with medication begins.

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.

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

0108_ADD_MEF_7.0_final.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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)*
IF a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. (Describe how and from whom their input was obtained.)
IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

Attention-deficit/hyperactivity disorder (ADHD) is a brain disorder marked by an ongoing pattern of inattention and/or hyperactivity-impulsivity that interferes with functioning or development. Medications can improve function, but proper monitoring is recommended. The intent of this measure is to ensure timely and continuous follow-up visits for children who are newly prescribed ADHD medication. The goal is to encourage monitoring of children for medication effectiveness, occurrence of side effects and adherence.

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 (4b) 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 the 10th, 25th, 50th, 75th and 90th percentile. Data are stratified by year and product line (i.e. commercial, Medicaid, HMO and PPO).

The following data demonstrate room for improvement among health plans.

These rates are extracted from HEDIS data collection and reflect the most recent years of measurement for this measure. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare 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 median eligible population for the measure across health plans.

INITIATION PHASE
Commercial
YEAR| MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 39.1% I 7.9% I 15.6% I 70.5% I 29.1% I 34.4% I 38.2% I 43.6% I 49.5%

2015 I 37.5% I 8.1% I 8.5% I 70.0% I 27.3% I 32.8% I 37.3% I 42.5% I 46.5%
2016 I 39.0% I 8.6% I 4.9% I 75.8% I 29.1% I 34.3% I 38.6% I 43.5% I 50.2%

Medicaid
YEAR| MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 39.6% I 11.3% I 7.4% I 64.9% I 21.8% I 32.6% I 41.1% I 47.0% I 53.0%
2015 I 40.1% I 10.8% I 9.0% I 69.7% I 25.6% I 32.9% I 40.8% I 49.1% I 54.0%
2016 I 42.2% I 11.0% I 0.0% I 77.7% I 28.8% I 34.2% I 42.2% I 49.6% I 55.5%

CONTINUATION AND MAINTENANCE PHASE
Commercial
YEAR| MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 45.9% I 9.2% I 23.1% I 70.8% I 35.1% I 39.7% I 45.3% I 51.2% I 57.8%
2015 I 44.7% I 8.8% I 20.6% I 78.3%I 35.1% I 39.5% I 43.7% I 49.8% I 56.0%
2016 I 46.8% I 9.3% I 22.7% I 81.8% I 35.6% I 40.7% I 46.4% I 52.3% I 57.3%

Medicaid
YEAR| MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 46.4% I 15.1% I 8.1% I 74.1% I 23.1% I 37.2% I 49.5% I 57.6% I 63.1%
2015 I 47.5% I 15.6% I 12.7% I 88.6% I 24.4% I 34.7% I 50.6% I 58.4% I 65.2%
2016 I 50.9% I 13.3% I 19.7% I 76.9% I 34.0% I 40.9% I 52.5% I 62.5% I 67.3%

Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

INITIATION PHASE
Commercial
YEAR| N PLANS | Mean Denominator Size per plan
2014 I 352 I 447.0
2015 I 340 I 438.9
2016 I 344 I 396.9

Medicaid
YEAR| N PLANS | Mean Denominator Size per plan
2014 I 154 I 1,121.3
2015 I 164 I 1,155.0
2016 I 180 I 1,160.2

CONTINUATION AND MAINTENANCE PHASE
Commercial
YEAR| N PLANS | Mean Denominator Size per plan
2014 I 228 I 185
2015 I 212 I 177
2016 I 210 I 163

Medicaid
YEAR| N PLANS | Mean Denominator Size per plan
2014 I 122 I 337
2015 I 143 I 319
2016 I 151 I 320

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 (4b) under Usability and Use.*
HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). NCQA does not currently collect performance data stratified by race, ethnicity, or language. Escarce et al. have described in detail the difficulty of collecting valid data on race, ethnicity and language at the health plan level (Escarce, 2011). 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. The HEDIS Health Plan Measure Set contains two measures that can assist with stratification to assess health care disparities. The 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
Studies suggest children from minority families experience decreased access to and utilization of health services for ADHD, even after controlling for poverty and health insurance status (Miller, Nigg, & Miller 2009). Although the prevalence of ADHD in minority children is believed to be equal to or even greater than the prevalence in non-minority children, studies indicate that rates of both diagnosis and treatment of ADHD are much lower among minority children compared to non-minority children. Specifically, children who are black, are raised in primarily non-English speaking households, have limited access to the health care system, and are poorer (Bailey et al. 2014; Morgan et al. 2014; Flores & Lin 2013; Froehlich et al. 2007)). Further studies indicate that, among children with ADHD, Hispanic and African American children were less often reported to use medication than white children (Bailey et al. 2014; Pastor & Reuben 2005). The NIMH Multisite Multimodal Treatment Study of Children with Attention-Deficit/Hyperactivity Disorder cited by American Academy of Child and Adolescent Psychiatry (AACAP) and American Academy of Pediatrics (AAP) indicates that certain disparities affected eight-year prospective follow-up. Participants lost to follow-up were "more often male, had younger mothers, had less educated parents, had lower parent income, and were more likely to have been on welfare at baseline" (Brooke et al. 2009). Studies suggest effective ADHD treatment in minority children may be affected by cultural norms surrounding ADHD. For example, some minority communities perceive that mental illness is a sign of personal weakness or that seeking treatment will jeopardize future employment or military service (Bailey et al. 2014). These perceptions lead to a lack of treatment seeking by these individuals and lack of appropriate screening (Price et al., 2013).

Bailey, R. K., Jaquez-Gutierrez, M. C., & Madhoo, M. 2014. "Sociocultural issues in african A\american and hispanic minorities seeking care for attention-deficit/hyperactivity disorder". The Primary Care Companion for CNS Disorders 16(4).

Brooke S.G. Molina Ph.D., Stephen P. Hinshaw Ph.D., James M. Swanson Ph.D., L. Eugene Arnold M.D., M.Ed., Benedetto Vitiello M.D., Peter S. Jensen M.D., Jeffery N. Epstein Ph.D., Betsy Hoza Ph.D., Lily Hechtman M.D., Howard B. Abikoff Ph.D., Glen R. Elliott Ph.D., M.D., Laurence L. Greenhill M.D., Jeffrey H. Newcorn M.D., Karen C. Wells Ph.D., Timothy Wigal Ph.D., Robert D. Gibbons Ph.D., Kwan Hur Ph.D. and Patricia R. Houck M.S. 2009. "The MTA at 8 Years: Prospective Follow-up of Children Treated for Combined-Type ADHD in a Multisite Study." Journal of the American Academy of Child and Adolescent Psychiatry 48(5):484-500.

Flores G. & H. Lin. 2013. "Trends in racial/ethnic disparities in medical and oral health, access to care, and use of services in US children: has anything changed over the years?" International Journal for Equity in Health 12:10.

Froehlich T.E., B.P. Lanphear, J.N. Epstein. 2007. "Prevalence, Recognition, and Treatment of Attention-Deficit/Hyperactivity Disorder in a National Sample of US Children." Archives of Pediatric and Adolescent Medicine 161(9):857-64.

Miller T.W., J.T. Nigg, R.L. Miller. 2009. "Attention deficit hyperactivity disorder in African American children: what can be concluded from the past ten years?" Clinical Psychological Review 29(1):77-86.

Morgan P.L., M.M. Hillemeier, G. Farkas, S. Maczuga. 2014. "Racial/ethnic disparities in ADHD diagnosis by kindergarten entry." Journal of Child Psychology and Psychiatry, and Allied Disciplines 55(8):905-13.

Pastor P.N. & C.A. Reuben. 2005. "Racial and ethnic differences in ADHD and LD in young school-age children: parental reports in the National Health Interview Survey." Public Health Reports 120(4): 383–392.

Price, J. H., Khubchandani, J., McKinney, M., & Braun, R. 2013. "Racial/ethnic disparities in chronic diseases of youths and access to health care in the United States". BioMed Research International.

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 : Attention Deficit Hyperactivity Disorder (ADHD)

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

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

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.)*
NA

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:** 0108_ADD_Value_Sets.xlsx

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 changes

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 children newly prescribed ADHD medication, those who had timely and continuous follow-up visits.

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).
RATE 1. INITIATION PHASE NUMERATOR
An outpatient, intensive outpatient or partial hospitalization follow-up visit with a practitioner with prescribing authority, within 30 days after the earliest prescription dispensing date for a new ADHD medication. Any of the following code combinations billed by a practitioner with prescribing authority meet criteria:
ADD Stand Alone Visits Value Set.
ADD Visits Group 1 Value Set with ADD POS Group 1 Value Set.
ADD Visits Group 2 Value Set with ADD POS Group 2 Value Set.
Note: Do not count a visit on the Index Prescription Start Date as the Initiation Phase visit.

RATE 2. CONTINUATION AND MAINTENANCE PHASE NUMERATOR
Children who are numerator compliant for Rate 1. Initiation Phase, AND have documentation of at least two follow-up visits with any practitioner from 31–300 days (9 months) after the earliest prescription dispensing date for a new ADHD medication.

One of the two visits (during days 31–300) may be a telephone visit (Telephone Visits Value Set) with any practitioner. Any of the following code combinations identify follow-up visits:
ADD Stand Alone Visits Value Set.
ADD Visits Group 1 Value Set with ADD POS Group 1 Value Set.
ADD Visits Group 2 Value Set with ADD POS Group 2 Value Set.
Telephone Visits Value Set.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Children 6-12 years of age newly prescribed ADHD 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).
RATE 1. INITIATION PHASE DENOMINATOR
Children age 6 as of March 1 of the measurement year; 12 years as of February 28 of the measurement year. who were dispensed a new ADHD medication during the 12-month Intake Period (Table ADD-A). Patients must have all of the following:(1) A 120-day (4-month) negative medication history on or before the Index Prescription Date. The Index Prescription Start Date is the dispensing date of the earliest ADHD prescription in the Intake Period with a Negative Medication History.

(2) Continuous enrollment for 120 days prior to the Index Prescription Start Date through 30 days after the Index Prescription Start Date.
(3) Exclude patients who had an acute inpatient encounter for mental health or chemical dependency during the 30 days after the Index Prescription Start Date. An acute inpatient encounter in combination with any of the following meet criteria:
A principal mental health diagnosis (Mental Health Diagnosis Value Set).
A principal diagnosis of chemical dependency (Chemical Dependency Value Set)

Due to the extensive volume of codes associated with identifying the denominator for this measure, we are attaching a separate file with code value sets. See code value sets located in question S.2b.

Table ADD-A: ADHD Medications
CNS stimulants: Amphetamine-dextroamphetamine, dexmethylphenidate, dextroamphetamine, lisdexamfetamine, methamphetamine, methylphenidate
Alpha-2 receptor agonists: Clonidine, guanfacine
Miscellaneous: Atomoxetine

RATE 2. CONTINUATION AND MAINTENANCE PHASE DENOMINATOR
Children who meet the eligible population criteria for Rate 1. Initiation Phase who have been continuously enrolled in the organization for 120 days (4 months) prior to the Index Prescription Start Date and 300 days (10 months) after the Index Prescription Start Date. Patients must have all of the following:
(1) The patient must have filled a sufficient number of prescriptions to provide continuous treatment for at least 210 days out of the 300-day period after the Index Prescription Start Date. The definition of "continuous medication treatment" allows gaps in medication treatment, up to a total of 90 days during the 300-day (10-month) period. (This period spans the Initiation Phase [1 month] and the C&M Phase [9 months].)
Gaps can include either washout period gaps to change medication or treatment gaps to refill the same medication.
Regardless of the number of gaps, the total gap days may be no more than 90. The organization should count any combination of gaps (e.g., one washout gap of 14 days and numerous weekend drug holidays).
(2) Exclude patients who had an acute inpatient encounter for mental health or chemical dependency during the 300 days (10 months) after the Index Prescription Start Date. An acute inpatient encounter in combination with any of the following meet criteria:
A principal mental health diagnosis (Mental Health Diagnosis Value Set).
A principal diagnosis of chemical dependency (Chemical Dependency Value Set).

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Children who had an acute inpatient encounter for mental health or chemical dependency following the Index Prescription Start Date
Children with a diagnosis of narcolepsy: Many of the medications used to identify patients for the denominator of this measure are also used to treat narcolepsy. Children with narcolepsy who are pulled into the denominator are then removed by the narcolepsy exclusion.
Children using hospice services during the measurement year. Children in hospice may not be able to receive the necessary follow-up care.

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 from the denominator for both rates, children who had an acute inpatient encounter for mental health or chemical dependency during the 30 days after the Index Prescription Start Date
Exclude from the denominator for both rates, children with a diagnosis of narcolepsy (Narcolepsy Value Set) any time during their history through December 31 of the measurement year
Exclude from the denominator for both rates 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 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).

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.)*
INITIATION PHASE: ELIGIBLE POPULATION
Step 1: Identify all children in the specified age range (Children 6-12 years of age: 6 as of March 1 of the measurement year; 12 years as of February 28 of the measurement year) who were dispensed an ADHD medication (Table ADD-A) during the 12-month Intake Period.
Step 2: Test for Negative Medication History. For each member identified in step 1, test each ADHD prescription for a Negative Medication History. The Index Prescription Start Date is the dispensing date of the earliest ADHD prescription in the Intake Period with a Negative Medication History.
Step 3: Calculate continuous enrollment. Patients must be continuously enrolled for 120 days (4 months) prior to the Index Prescription Start Date through 30 days after the Index Prescription Start Date.
Step 4: Exclude patients who had an acute inpatient encounter for mental health or chemical dependency during the 30 days after the Index Prescription Start Date. An acute inpatient encounter (Acute Inpatient Value Set) in combination with any of the following meet criteria: A principal mental health diagnosis (Mental Health Diagnosis Value Set) AND/OR A principal diagnosis of chemical dependency (Chemical Dependency Value Set).
Step 5: Determine the number of patients in the eligible population with an outpatient, intensive outpatient or partial hospitalization follow-up visit with a practitioner with prescribing authority, within 30 days after the Index Prescription Start Date. Any of the following code combinations billed by a practitioner with prescribing authority meet criteria:
ADD Stand Alone Visits Value Set.
ADD Visits Group 1 Value Set with ADD POS Group 1 Value Set.
ADD Visits Group 2 Value Set with ADD POS Group 2 Value Set.
Note: Do not count a visit on the Index Prescription Start Date as the Initiation Phase visit.
Step 6: Calculate a rate (number of children receiving a follow-up visit with a prescriber within 30 days of the Index Prescription Start Date).

CONTINUATION AND MAINTENANCE PHASE: ELIGIBLE POPULATION
Step 1: Identify all patients who meet the eligible population criteria for Rate 1—Initiation Phase.
Step 2: Calculate continuous enrollment. Patients must be continuously enrolled in the organization for 120 days (4 months) prior to the Index Prescription Start Date and 300 days (10 months) after the Index Prescription Start Date.
Step 3: Calculate the continuous medication treatment. Using the patients in step 2, determine if the member filled a sufficient number of prescriptions to provide continuous treatment for at least 210 days out of the 300-day period after the Index Prescription Start Date. The definition of “continuous medication treatment” allows gaps in medication treatment, up to a total of 90 days during the 300-day (10-month) period. (This period spans the Initiation Phase [1 month] and the C&M Phase [9

months].) Gaps can include either washout period gaps to change medication or treatment gaps to refill the same medication. Regardless of the number of gaps, the total gap days may be no more than 90. The organization should count any combination of gaps (e.g., one washout gap of 14 days and numerous weekend drug holidays).
Step 4: Exclude patients who had an acute inpatient encounter for mental health or chemical dependency during the 300 days (10 months) after the Index Prescription Start Date. An acute inpatient encounter in combination with any of the following meet criteria:
A principal mental health diagnosis (Mental Health Diagnosis Value Set).
A principal diagnosis of chemical dependency (Chemical Dependency Value Set).
Step 5: Identify all patients in the eligible population who meet the following criteria:
(1) Numerator compliant for Rate 1—Initiation Phase, and
(2) At least two follow-up visits from 31–300 days (9 months) after the Index Prescription Start Date with any practitioner.
One of the two visits (during days 31–300) may be a telephone visit (Telephone Visits Value Set) with any practitioner. Any of the following code combinations identify follow-up visits:
ADD Stand Alone Visits Value Set.
ADD Visits Group 1 Value Set with ADD POS Group 1 Value Set.
ADD Visits Group 2 Value Set with ADD POS Group 2 Value Set.
Telephone Visits Value Set.
Step 6: Calculate a rate (number of children receiving two follow-up visits with any practitioner from 31-300 days after the Index Prescription Start Date).

ADDITIONAL EXCLUSION:
Exclude from the denominator for both rates, patients with a diagnosis of narcolepsy (Narcolepsy Value Set) any time during their history through December 31 of the measurement year

NOTE
(1) Patients who have multiple overlapping prescriptions should count the overlap days once toward the days supply (whether the overlap is for the same drug or for a different drug).
(2) Organizations may have different methods for billing intensive outpatient encounters and partial hospitalizations. Some methods may be comparable to outpatient billing, with separate claims for each date of service; others may be comparable to inpatient billing, with an admission date, a discharge date and units of service. Organizations whose billing methods are comparable to inpatient billing may count each unit of service as an individual visit. The unit of service must have occurred during the period required for the rate (e.g., within 30 days after or from 31–300 days after the Index Prescription Start Date).

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

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.)*
IF a PRO-PM, 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 (Only), Pharmacy

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 is collected.)*
IF a PRO-PM, identify the specific PROM(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 Management Organizations and Preferred Provider Organizations 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

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Clinician Office/Clinic
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
0108_ADD_MTF_7.0_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. (Do not remove prior testing information – include date of new information in red.)
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. (Do not remove prior testing information – include date of new information in red.)
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0108
Measure Title: Follow-Up Care for Children Prescribed ADHD Medication
Date of Submission: 12/2/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for*

measure implementation. ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

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

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

1.3. What are the dates of the data used in testing? Click here to enter date range
2011-2012
2014-2016

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

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

1.5. How many and which <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*)

<u>2016 Update: MEASURE SCORE RELIABILITY TESTING</u>

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

<u>2013 Submission</u>

MEASURE SCORE RELIABILITY TESTING

The measure score reliability was calculated from HEDIS data that included 177 Commercial HMO health plans, 180 Commercial PPO health plans, and 140 Medicaid health plans for the Initiation Phase and 99 Commercial HMO health plans, 135 Commercial PPO health plans, and 116 Medicaid health plans for the Continuation and Maintenance Phase. The sample included all Commercial HMO, Commercial PPO and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

CONSTRUCT VALIDITY TESTING

Construct validity was calculated from HEDIS data that included 357 Commercial health plans for the Initiation Phase and 234 Commercial health plans for the Continuation and Maintenance Phase. The sample included all Commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

SYSTEMATIC EVALUATION OF FACE VALIDITY

This measure was tested for face validity with four panels of experts. See Additional Information: Ad.1.

Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel:

The Behavioral Health Measurement Advisory Panel included 12 experts in behavioral health, psychiatry, child psychology, public health, quality measurement, behavioral health care, education, and serious mental illness, including representation by consumers, health plans, health care providers and policymakers. The Technical Measurement Advisory Panel includes 14 members, including representation by health plans, methodologists, clinicians and HEDIS auditors. NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the HEDIS measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is composed of 15 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement. The HEDIS Expert Coding Panel includes 11 members, including representation by health plans, hospital associations, and advisory groups.

VALIDITY ASSESSMENT IN FIELD TESTING

The field test data, collected when the measure was first established in 2004, included 6 health plans that provided member-level administrative data to NCQA for the field test.

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*)

<u>2016 Update: MEASURE SCORE RELIABILITY TESTING</u>

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

INITIATION PHASE

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

CONTINUATION AND MAINTENANCE PHASE

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

<u>2013 Submission</u>

MEASURE SCORE RELIABILITY TESTING

Patient Sample for measure score reliability testing: In 2013, HEDIS measures covered 108.1 million commercial health plan beneficiaries, 20.3 million Medicaid beneficiaries, and 14.0 million Medicare beneficiaries. This measure applies to commercial and Medicaid plans. Data are summarized at the health plan level and stratified by product line. Below is a description of the sample, including number of health plans included and the median eligible population for the measure across health plans.

INITIATION PHASE

Product Line	Number of Plans	Median number of eligible patients per plan
Commercial HMO	177	148
Commercial PPO	180	230
Medicaid	140	594

CONTINUATION AND MAINTENANCE PHASE

Product Line	Number of Plans	Median number of eligible patients per plan
Commercial HMO	99	82
Commercial PPO	135	90
Medicaid	116	183

CONSTRUCT VALIDITY TESTING

Beneficiary Sample for Construct Validity Testing: In 2013, HEDIS measures covered 108.1 million commercial health plan beneficiaries, 20.3 million Medicaid beneficiaries, and 14.0 million Medicare beneficiaries. Data is summarized at the health plan level. Construct validity was calculated from HEDIS data that 177 Commercial HMO health plans, 180 Commercial PPO health plans, and 140 Medicaid health plans for the Initiation Phase and 99 Commercial HMO health plans, 135 Commercial PPO health plans, and 116 Medicaid health plans for the Continuation and Maintenance Phase. The sample included all Commercial and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

VALIDITY ASSESSMENT IN FIELD TESTING

Patient Sample for Field Test Validity Assessment: The field test data, collected when the measure was first established in 2004, included 6 health plans that provided member-level administrative data to NCQA. Plans' included both commercial and Medicaid product lines, with representative membership ranging in size from 3,500 to 1 million. The service areas of the participating plans were also extensive, providing comprehensive coverage to select states or regions from a variety of geographic areas within the United States. Participating plans were asked to submit enrollment, encounter and medication data to NCQA, who then performed the actual calculations of the measure rates.

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.

2016 Update: MEASURE SCORE RELIABILITY TESTING

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

2013 Submission

MEASURE SCORE RELIABILITY TESTING

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

SYSTEMATIC EVALUATION OF FACE VALIDITY

Validity was demonstrated through a systematic assessment of face validity as well as by assessing construct validity by examining correlations with other measures. Per NQF instructions we have described the composition of the technical expert panel which assessed face validity in the data sample questions above. To assess construct validity the correlation was examined between this measure and the Children and Adolescent' Access to Primary Care Practitioners HEDIS measure, which also includes the entire HEDIS data sample.

VALIDITY ASSESSMENT IN FIELD TESTING

When validity was assessed during initial field testing in 2004, medical record documentation was used as a gold standard to assess whether the administrative data sources work well for identifying patients with a new ADHD medication episode.

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

2016 Update: Measure performance was asssessed by commercial, Medicaid and Medicare plan type.

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*)

2016 Update: METHOD FOR MEASURE SCORE RELIABILITY TESTING

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

2013 Submission

METHOD FOR BETA-BINOMIAL RELIABILITY TESTING

The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the *signal*. The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents *noise*. The reliability of the measure is represented as the ratio of signal to noise.

- A score of 0 indicates none of the variation (signal) is attributable to the plan
- A score of 1.0 indicates all of the variation (signal) is attributable to the plan
- A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY

The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan's variation for the average plan's variation. The reliability for some plans may be more or less than the overall reliability across plans.

Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corp. TR-653-NCQA, 2009

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*)

2016 Update: MEASURE SCORE RELIABILITY

Beta-Binomial Statistic For Each Measure Rate: Mean Reliability

Rate	*Commercial*	*Medicaid*
Initiation Phase	0.90	0.98
Continuation and Maintenance Phase	0.75	0.95

2013 Submission

MEASURE LEVEL RELIABILITY

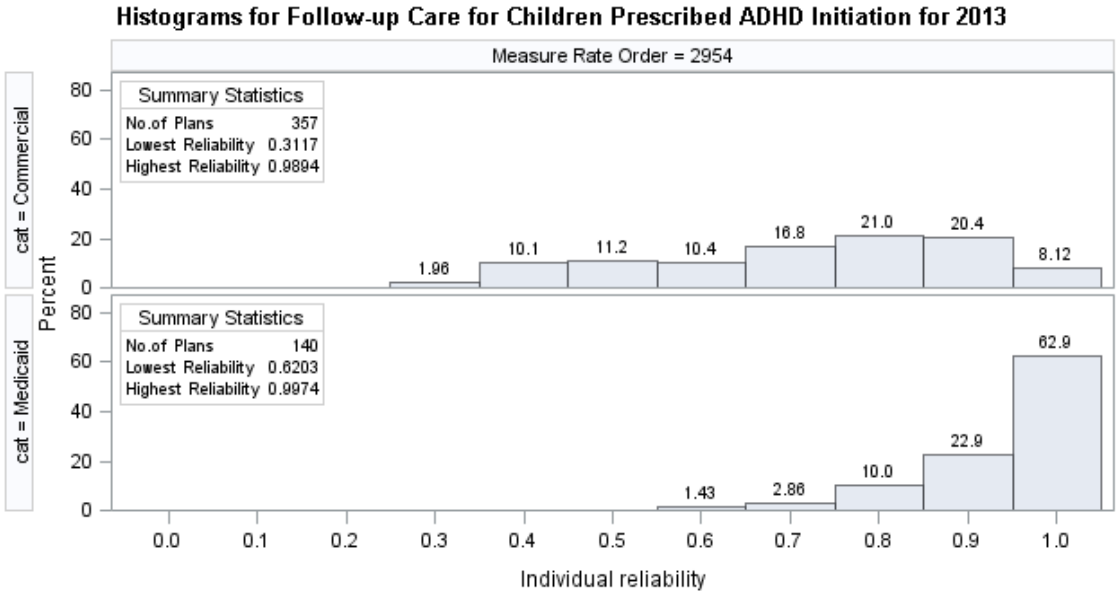
NCQA pools data reported by health plans according to product line. The mean reliability for the Initiation Phase as per the beta binomial model was 0.71 for Commercial health plans and 0.93 for Medicaid. The mean reliability for the Continuation and Maintenance Phase was 0.66 for Commercial health plans and 0.92 for Medicaid.

Beta-Binomial Statistic For Each Measure Rate

Rate	Commercial			Medicaid		
	Avg	SD	10-90th	Avg	SD	10-90th
Initiation Phase	0.7	0.2	0.4-0.9	0.9	0.1	0.8-0.99
Continuation and Maintenance Phase	0.7	0.2	0.4-0.9	0.9	0.1	0.8-0.98

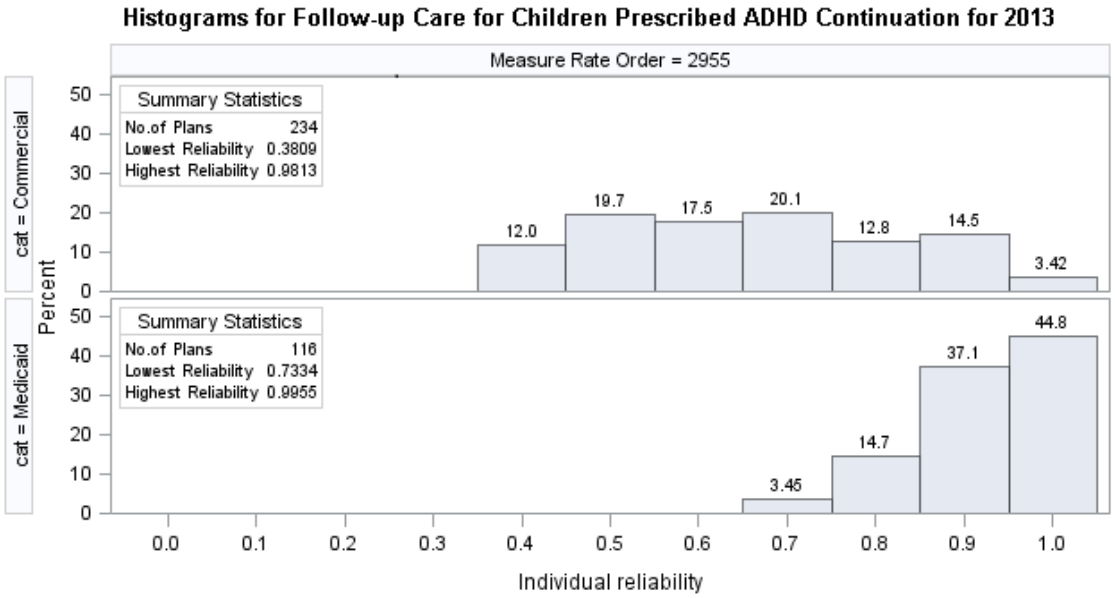
HEALTH-PLAN LEVEL RELIABILITY

The histograms show the distribution of the individual reliability values for each product line.



INITIATION PHASE

CONTINUATION AND MAINTENANCE 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?*)

2016 Update: INTERPRETATION OF RESULTS FOR MEASURE SCORE RELIABILITY TESTING

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

2013 Submission

Among both Medicaid and commercial plans, results indicate both the Initiation Phase and Continuation and Maintenance Phase rates within this measure have a good signal to noise ratio, thus having sufficient signal strength to discriminate performance between accountable entities.

At the plan level, the vast majority of Medicaid plans met or exceeded the 0.7 threshold. Commercial plans showed more distribution, though at least half of plans met the 0.7 threshold for both rates.

2b2. VALIDITY TESTING

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

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

☐ **Performance measure score**

☒ **Empirical validity testing**

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

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

METHOD OF ASSESSING FACE VALIDITY

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 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 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 publically 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 next year's HEDIS Volume 2.

METHOD OF TESTING CONSTRUCT VALIDITY

We tested for construct validity by exploring whether the rates of this measure were correlated with another HEDIS measure, Children and Adolescents Access to Primary Care Practitioners. We hypothesized the two measures would be positively correlated, as organizations that perform well on providing follow-up care for children on an ADHD medication should also perform well on providing access to primary care practitioners. We used a Pearson correlation test, which estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +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 FOR VALIDITY ASSESSMENT IN FIELD TESTING

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 for the entire eligible population. Given the eligible population, plans also reviewed a sample of 150 patients' medical records specifically located at the provider responsible for triggering the Index Start Date. If the medical record was not available with the Index provider, then plans sought the patient's chart at the primary care physician's office. The reason for including certain information from both administrative sources and medical records was to verify the completeness and accuracy of the administrative data. These field test methods were designed to answer three questions with respect to validity:

1. Is administrative data available for identifying eligible patients? Can it accurately identify eligible patients relative to the medical record? Should the denominator require a diagnosis of ADHD medication in addition to a new medication episode? Should the new medication episode allow off-label ADHD medications?
2. Is administrative data concerning ADHD follow-up visits (numerator) available? Is it accurate relative to the medical record? Should the numerator allow any type of follow-up visit or follow-up visits that contain an ADHD-specific code?
3. Does performance vary depending upon the denominator specification or the numerator specification?

ICD-10 CONVERSION

The below steps describe our methods to convert this measure to ICD-10 in order to develop a new code set fully consistent with the intent of the measure.

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

Tools Used to Identify/Map to ICD-10

All tools used for mapping/code identification from CMS ICD-10 website (http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html).
GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

Expert Participation

The NCQA HEDIS Expert Coding Panel and NCQA's Behavioral Health Measurement Advisory Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

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

RESULTS OF FACE VALIDITY ASSESSMENT

Step 1: This measure was developed to address insufficient follow-up care for children with a prescription for an ADHD medication in 2004. NCQA and the Behavioral Health MAP worked together to assess the most appropriate tools for assessing initial and continuous follow-up care for children and adolescents on an ADHD medication.

Step 2: The measure was written and field-tested from 2004-2005. After reviewing field test results, The CPM recommended to send the measure to public comment with a majority vote in 2005.

Step 3: The measure was released for Public Comment in 2005 prior to publication in HEDIS. We received and responded to 133 comments on this measure. The majority of comments supported the measure. The CPM recommended moving this measure to first year data collection by a majority vote.

Step 4: The measure was introduced in HEDIS 2006. 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 to public reporting with a majority vote.
Conclusion: The measure was deemed to have the desirable attributes of a HEDIS measure in 2005 (relevance, scientific soundness, and feasibility).

RESULTS OF CONSTRUCT VALIDITY ASSESSMENT

The results in Table 1 showed that the ADHD measure was significantly and positively correlated with the *Children and Adolescents Access to Primary Care Practitioners measure* among commercial plans. The level of correlation among these measures was moderate.

Table 1. Pearson Correlation Coefficients between *Follow-up Care for Children Prescribed ADHD Medication* and *Child and Adolescent Access to Primary Care Practitioners*: Commercial Plans, 2013

	Access to Primary Care Practitioners
ADHD Follow-up: Initiation Rate	0.4
ADHD Follow-up: Continuation & Maintenance Rate	0.4

All correlations are significant at p<.05

AMONG MEDICAID PLANS, THERE WERE NO CORRELATIONS ABOVE THE 0.3 THRESHOLD.RESULTS FROM FIELD TEST VALIDITY ASSESSMENT

Overall, the New ADHD Medication Episode (defined as patients with an ADHD prescription and a negative medication history > 120 days) was confirmed for 50.5% of the patients for whom a medical record could be located. In testing to see whether an ADHD diagnosis should also be included as a denominator requirement, the medical record validation for patients with an ADHD medication and an ADHD diagnosis on average was 43.8%, while 39% of the time the diagnosis was documented only in the medical record. One likely reason for these lower than expected rates is that the health plans indicated they were not always able to retrieve the entire medical record primarily due to resource constraints. Another reason is that pharmacy claims data is typically considered accurate and timely and the medical record may lack prescription information in the medical records reviewed.

In addition, among those confirmed as having a new medication episode, a diagnosis of ADHD was confirmed in the administrative or medical record data 85% of the time. This finding further supports our assessment that ADHD diagnosis should not be required as part of the denominator.

Finally, the measure excludes off-label medications since 96.8% of children newly prescribed an ADHD-specific medication had a confirmed diagnosis of ADHD in either medical record or administrative data. Looking separately at off-label ADHD medications, the confirmation of ADHD diagnosis in the medical record or administrative data dropped to 42%.

The medical record validation of the measure's two indicators, Initiation Phase and Continuation & Maintenance Phase, showed that the administrative data captured the follow-up visits found in the medical record:

-Initiation: 94% of the follow-up visit data in the medical record were appropriately submitted to the health plans as administrative data.

-Continuation & Maintenance: 70% of the follow-up visit data in the medical record were submitted to the health plan as administrative data.

Because the only specification difference between compliance for the Initiation visit and Continuation & Maintenance follow-up visits are when the visits take place with respect to the index prescription event, this drop in administrative data matching the medical record data may primarily be due to data submission lags. The administrative data necessary to accurately report compliance for these indicators is considered valid and complete.

During field testing, different denominator specifications ("requiring an ADHD diagnosis" vs. "not requiring an ADHD diagnosis") and different numerator specifications (requiring "any visit," "a visit with a principal ADHD code," or a "visit with a principal or secondary ADHD code") were tested to assess comparability across plans, as well as rates. Based on the field test data below, the current measure does not require an ADHD diagnosis for the denominator and allows "any visit" for the numerator.

These are performance rates by numerator requirement if an ADHD diagnosis is required for the denominator:
-Initiation: "Any visit" numerator rate=47.9%; ADHD-specific visit coded primarily=29.7%; ADHD-specific visit coded primarily or secondarily=31.4%
-Continuation & Maintenance: "Any visit" numerator rate=45.0%; ADHD-specific visit coded primarily=22.2%; ADHD-specific visit coded primarily or secondarily=23.4%.

These are performance rates by numerator requirement if an ADHD diagnosis is not required for the denominator:

-Initiation: "Any visit" numerator rate=42.3%; ADHD-specific visit coded primarily=17.5%; ADHD-specific visit coded primarily or secondarily=19.3%
-Continuation & Maintenance: "Any visit" numerator rate=40.0%; ADHD-specific visit coded primarily=12.7%; ADHD-specific visit coded primarily or secondarily=13.7%

ICD-10 CONVERSION
Summary of Stakeholder Comments Received
NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. Comments received helped to ensure we were mapping the codes correctly.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
INTERPRETATION OF SYSTEMATIC ASSESSMENT OF FACE VALIDITY
Our advisory panels agreed that the measures as specified will accurately differentiate quality across providers. The measure had sufficient face validity.

INTERPRETATION OF CONSTRUCT VALIDITY TESTING
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. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample 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. The results confirmed the hypothesis that health plans with good rates of follow-up care for children with a prescription for an ADHD medication also have better performance on the Children and Adolescents Access to Primary Care Practitioners measure. These results indicate that the follow-up care for children with a prescription for ADHD medication measure is a valid measure of a plan's quality at providing children with appropriate care.

FIELD TEST VALIDITY ASSESSMENT: *See section 2a.2 above.*
The measure is feasible for administrative-only implementation with respect to the eligible population. Using medical record documentation as a gold standard, the administrative data sources appear to work well for identifying patients with a new ADHD medication episode (defined as patients with an ADHD prescription and a negative medication history > 120 days).

_____

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

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
The exclusion for this measure is based on clearly specified ICD-9/10 CM Diagnosis codes for narcolepsy. While these codes have not been tested in the context of this measure for validity, they are widely used across practitioners and considered to be valid. This measure does not allow for exclusions for patient refusal, provider refusal, or un-specified exclusions.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis.*
*Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is*

transparent, e.g., scores with and without exclusion)
N/A

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

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

2b4.1.1 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.

2b4.2. 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.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

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

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
2b4.9. Results of Risk Stratification Analysis:
2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

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 .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However the method can be used for comparison of any two measured entities.

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

2016 Update: ABILITY TO IDENTIFY STATISTICALLY SIGNIFICANT/MEANINGFUL DIFFERENCES

HEDIS 2016 Variation in Performance across Health Plans- Commercial

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

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

IQR: Interquartile range

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

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

Boxplot Graph for Commercial ADD Initiation Rate from 2014-2016

Summary Statistics			
No.of Plans	352.00	340.00	344.00
Average	39.101	37.448	39.004
Lowest Rate	15.625	8.4746	4.8780
Highest Rate	70.492	70.000	75.781

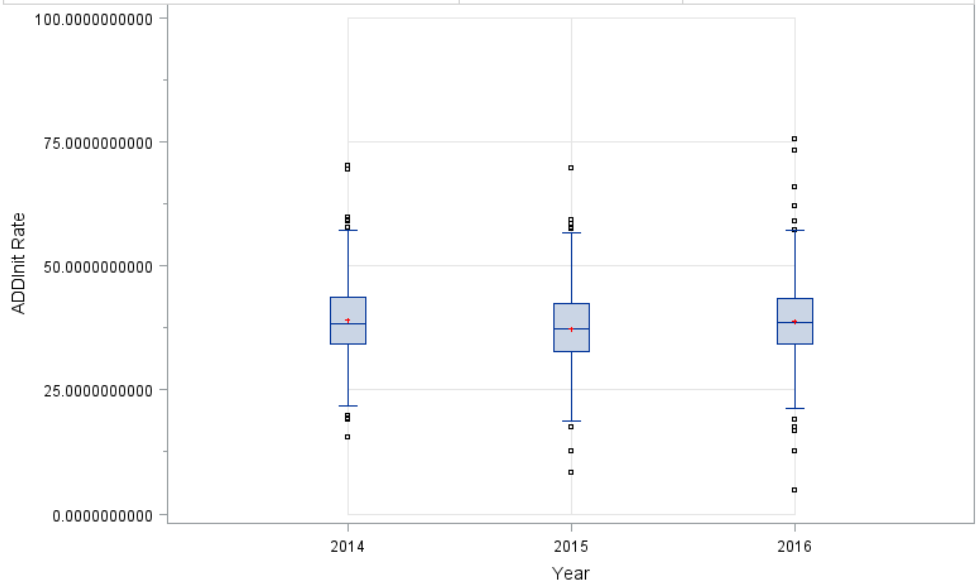


Figure 1b. *Follow-up Care for Children Prescribed ADHD Medication – C&M Phase*: Commercial Plans 2014-2016

Boxplot Graph for Commercial ADD Continuation Rate from 2014-2016

Summary Statistics			
No.of Plans	229.00	212.00	210.00
Average	45.944	44.737	46.750
Lowest Rate	23.077	20.588	22.727
Highest Rate	70.769	78.261	81.818

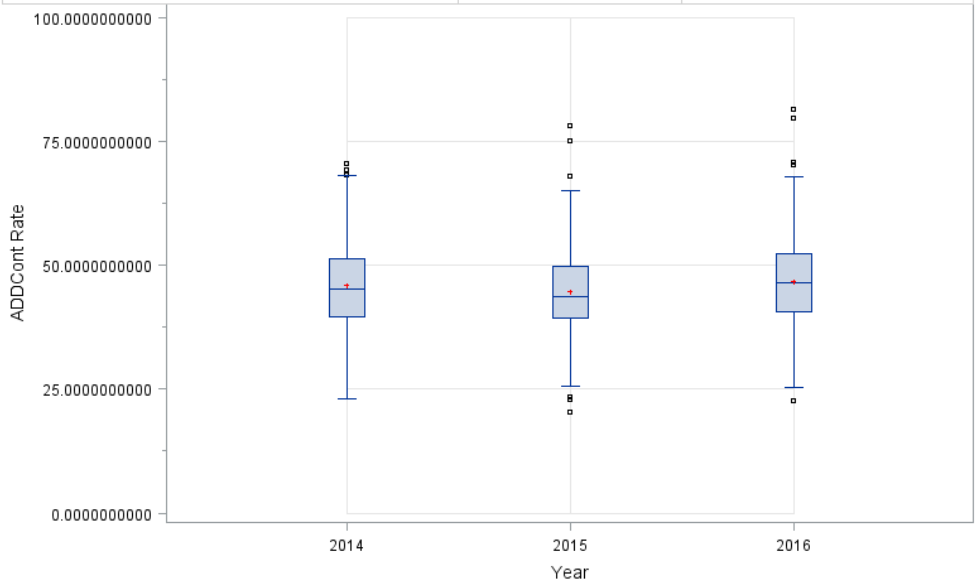


Figure 2a. *Follow-up Care for Children Prescribed ADHD Medication - Initiation Phase*: Medicaid Plans 2014-2016

Boxplot Graph for Medicaid ADD Initiation Rate from 2014-2016

Summary Statistics			
No.of Plans	154.00	164.00	180.00
Average	39.600	40.136	42.198
Lowest Rate	7.3903	8.9494	0.0000
Highest Rate	64.859	69.725	77.647

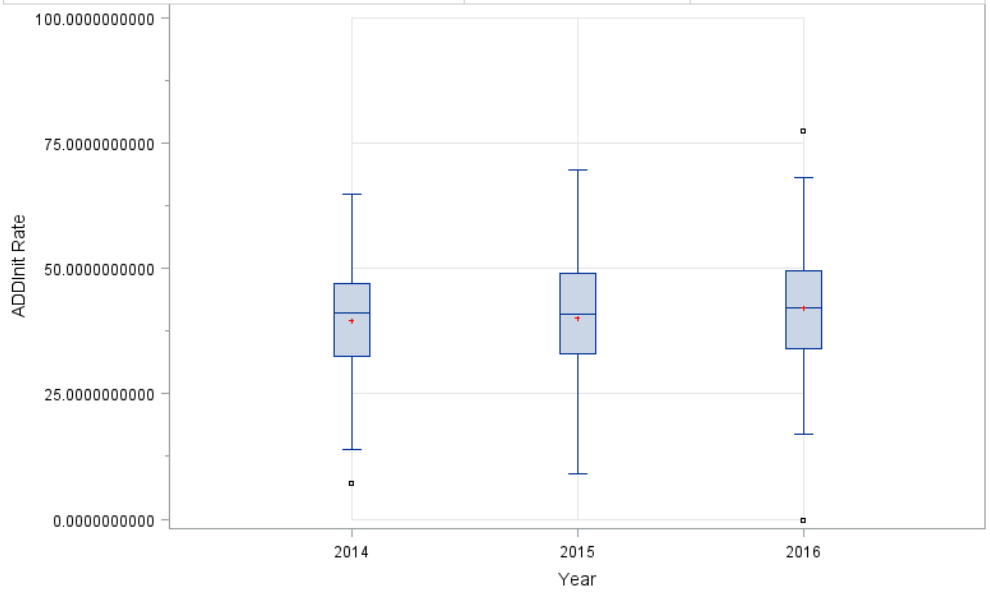
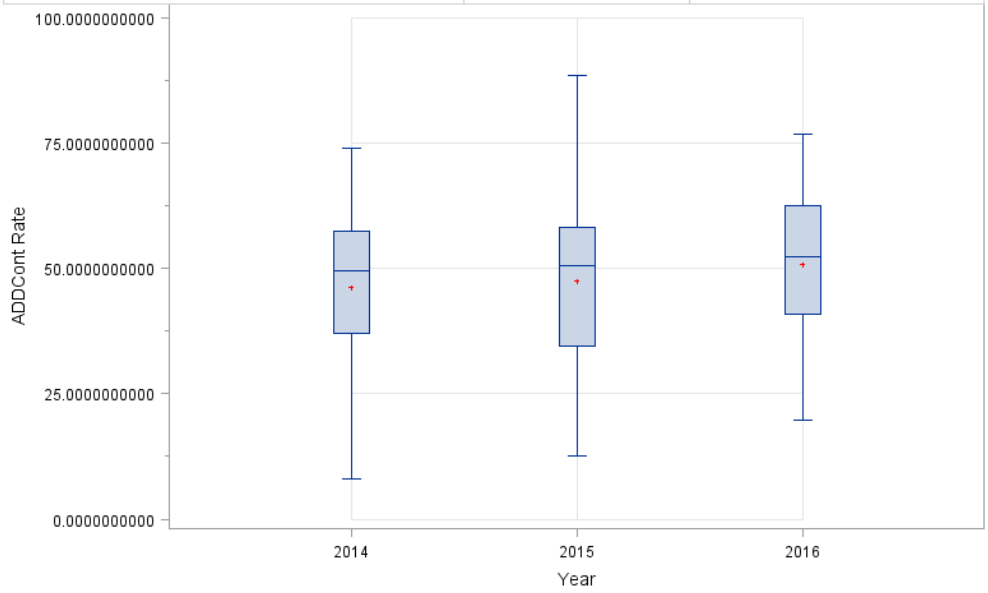


Figure 2b. *Follow-up Care for Children Prescribed ADHD Medication – C&M Phase*: Medicaid Plans 2014-2016

Boxplot Graph for Medicaid ADD Continuation Rate from 2014-2016

Summary Statistics			
No.of Plans	122.00	143.00	151.00
Average	46.352	47.517	50.931
Lowest Rate	8.1301	12.652	19.697
Highest Rate	74.109	88.571	76.923



2013 Submission

HEDIS 2013 Variation in Performance across Health Plans

	Rate	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Commercial	Initiation	440	38.4	7.7	29.3	34.1	37.5	43.0	48.9	8.9	0.001
HMO & PPO	C&M	181	45.3	9.4	34.5	39.5	45.0	50.2	57.1	10.7	0.023
Medicaid	Initiation	1,005	39.1	10.9	24.2	31.6	39.8	45.9	51.8	14.4	<0.001
HMO & PPO	C&M	289	45.2	14.8	25.0	34.7	46.7	55.9	63.8	21.2	<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
The below box plots show the distribution of performance rates for commercial and Medicaid plans

Figure 1a. ***Follow-up Care for Children Prescribed ADHD Medication - Initiation Phase*****: Commercial Plans 2011-2013**

No.of Plans	351.00	339.00	357.00
Average	38.478	39.372	38.372
Lowest Rate	8.3333	12.941	15.517
Highest Rate	60.870	62.025	69.792

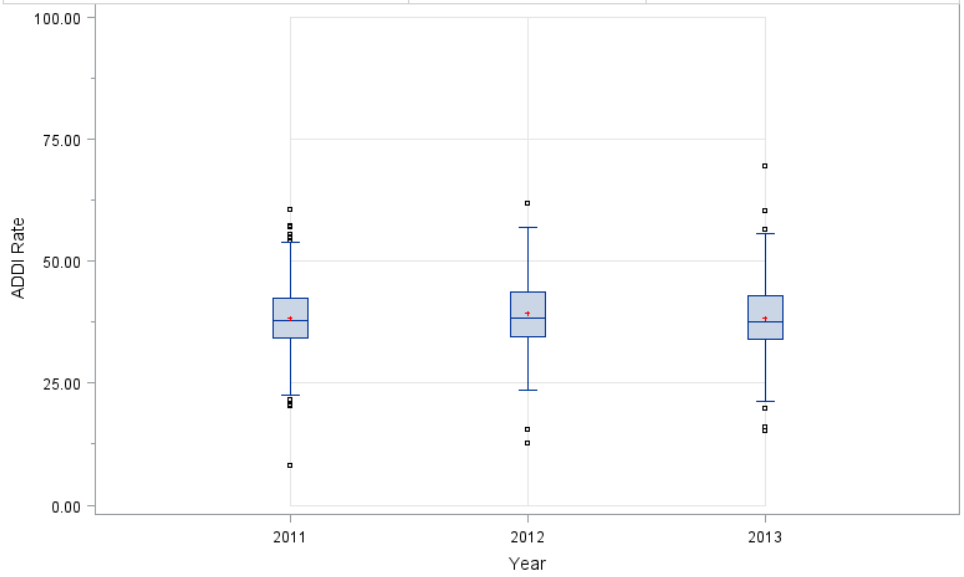


Figure 1b. ***Follow-up Care for Children Prescribed ADHD Medication – C&M Phase*****: Commercial Plans 2011-2013**

No.of Plans	227.00	220.00	234.00
Average	43.347	44.579	45.263
Lowest Rate	20.000	17.308	24.242
Highest Rate	88.571	88.559	95.000

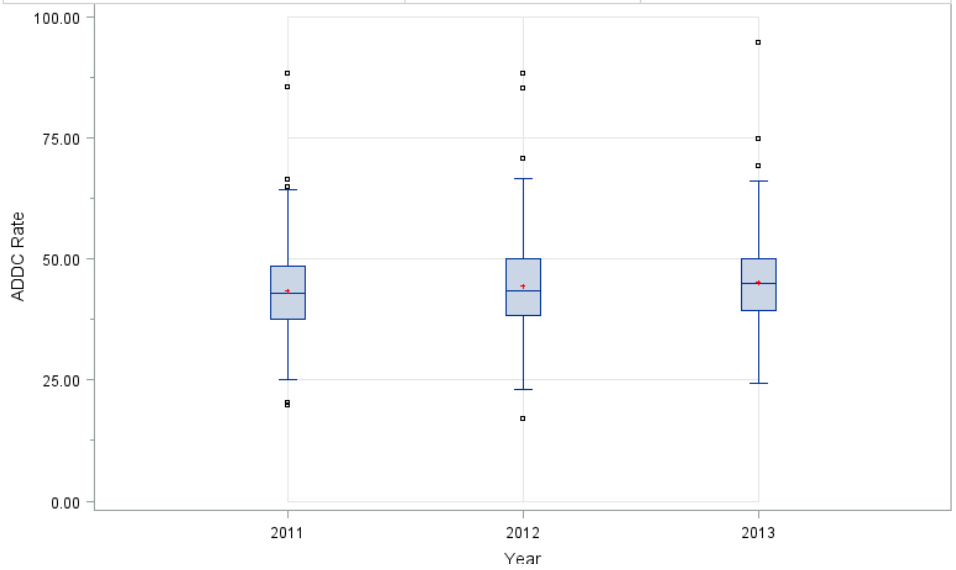
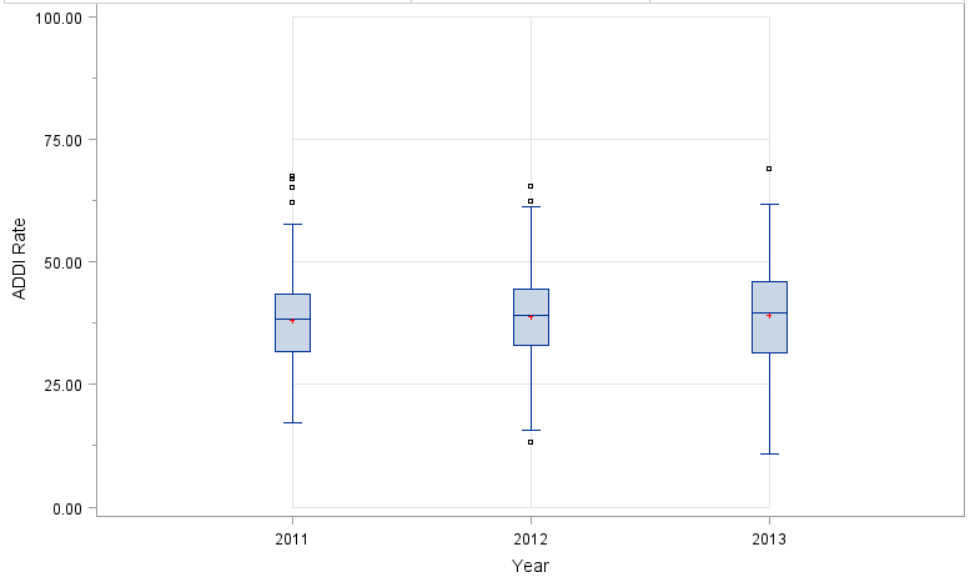


Figure 2a. *Follow-up Care for Children Prescribed ADHD Medication - Initiation Phase*: Medicaid Plans 2011-2013

No.of Plans	129.00	138.00	140.00
Average	38.103	38.832	39.103
Lowest Rate	17.330	13.493	10.886
Highest Rate	67.742	65.714	69.091



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

2016 Update: INTERPRETATION OF ABILITY TO IDENTIFY STATISTICALLY SIGNIFICANT/MEANINGFUL DIFFERENCES

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

2013 Submission

The results above indicate there is a 9-21% gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the Medicaid health plans which show a 14.4-21.2 percentage point gap between 25th and 75th percentile plans. This gap represents on average 145 children in the Initiation Phase and 61 children in the Continuation and Maintenance Phase in high performing Medicaid plans compared to low performing plans (estimated from average health plan eligible population).

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

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

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing***

performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

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

_____
2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased 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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.

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 organization´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 to questions in order to prevent possible errors or inconsistencies in the implementation of the measure. 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 usability 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 Health Plan Ranking http://reportcard.ncqa.org/plan/external/plansearch.aspx Health Plan Ranking http://reportcard.ncqa.org/plan/external/plansearch.aspx Payment Program Physician Quality Reporting System http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/EducationalResources.html?gclid=COjY783v278CFQto7AodHl8A0w CMS EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms/

	Physician Value-Based Payment Modifier (VBM) https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html Physician Quality Reporting System http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/EducationalResources.html?gclid=COjY783v278CFQto7AodHl8A0w CMS EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms/ Physician Value-Based Payment Modifier (VBM) https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html Regulatory and Accreditation Programs Accreditation: http://www.ncqa.org/tabid/123/Default.aspx Accountable Care Organization Accreditation: http://www.ncqa.org/Programs/OtherPrograms/acomeasuresPilotProject.aspx Accreditation: http://www.ncqa.org/tabid/123/Default.aspx Accountable Care Organization Accreditation: http://www.ncqa.org/Programs/OtherPrograms/acomeasuresPilotProject.aspx Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

4a.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

HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2012, a total of 455 Medicare Advantage health plans, 404 commercial health plans and 136 Medicaid health plans across 50 states were included in the rankings.

STATE OF HEALTH CARE ANNUAL REPORT: This measure is publically 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 2012 the report included measures on 11.5 million Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

MEDICAID CHILD CORE SET: This measure is included in the Medicaid Child Core Set which is a set of children's health care quality measures developed as part of the Children's Health Insurance Program (CHIP) Reauthorization Act for voluntary use by State Medicaid and CHIP programs. The data collected with these measures will help CMS to better understand the quality of health care children receive through Medicaid and CHIP and assist CMS and states in moving toward a national system for quality measurement, reporting, and improvement. As per the CHIPRA legislation, state data derived from the core measures will become part of the Secretary's annual report on the quality of care for children in Medicaid and CHIP. The Secretary's annual

report summarizes state-specific and national measurement information on the quality of health care furnished to children enrolled in Medicaid and CHIP.

PHYSICIAN QUALITY REPORTING SYSTEM: This measure is used in the Physician Quality Reporting System (PQRS) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals. Eligible professionals who satisfactorily report data on quality measures for covered Physician Fee Schedule services furnished to Medicare Part B beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer) receive these payment incentives and adjustments.

CMS EHR INCENTIVE PROGRAM: This measure is used in the CMS Electronic Health Record (EHR) Incentive Program, which provides incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology.

NCQA HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries. [REPLACE or ADD as appropriate, 336 commercial health plans covering 87 million lives; 77 Medicaid health plans covering 9.1 million lives.] Health plans are scored based on performance compared to benchmarks.

NCQA ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, 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.

PHYSICIAN VALUE-BASED PAYMENT MODIFIER (VBM): This measure is used in the Physician Value-Based Modifier program, which provides differential payment to a physician or group of physicians under the Medicare Physician Fee Schedule(PFS). 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.

QUALIFIED HEALTH PLAN (QHP) QUALITY RATING SYSTEM (QRS): This measure is used in the Qualified Health Plan (QHP) Quality Rating System, which provides comparable information to consumers about the quality of health care services and QHP enrollee experience offered in the Marketplaces.

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

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

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.

4b. **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 2014 to 2016, performance rates have been generally stable or shown slight improvement across commercial and Medicaid plans. For commercial plans, performance on average was 39% and 46% for the Initiation Phase and Continuation and Maintenance Phase rates, respectively. For Medicaid plans, the average performance increased two percentage points for the Initiation Phase rate and five percentage points for the Continuation and Maintenance Phase rate. However, across both commercial and Medicaid plans, there continues to be fairly large variation between the 10th and 90th percentiles, suggesting room for improvement. For example, among commercial plans, the 2016 rate of children who had documentation of a timely follow-up visit ranged from 29% for plans in the 10th percentile to 50% among plans in the 90th percentile.

4c. Unintended Consequences

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

4c.1. **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.

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

4d1.1. **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.

4d1.2. **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, as described in Section 3c1.

4d2.1. **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.

4d2.2. **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 information in claims meets the measure intent and questions about the supporting guidelines for the measure. NCQA responded to all questions to ensure consistent implementation of the specifications.

4d2.3. **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 Medicaid Child Core Set, CMS EHR Incentive Program and CMS Physician Quality Reporting Initiative.

4d.3. **Describe how the feedback described in 4d.2 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.

5. Comparison to Related or Competing Measures

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

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

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

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

5a. Harmonization 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?

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

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
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.
BEHAVIORAL HEALTH MEASUREMENT ADVISORY PANEL
Michael Schoenbaum, Ph.D., Senior Advisor for Mental Health Services, Epidemiology and Economics, National Institute of Mental Health
Frank A. Ghinassi, Ph.D., Vice President, Quality and Performance Improvement, Western Psychiatric Institute and Clinic and UPMC Behavioral Health Network, University of Pittsburgh Medical Center, Assistant Professor in Psychiatry University of Pittsburgh School of Medicine
Charlotte Mullican, B.S.W., M.P.H. Sr. Advisor for Mental Health Research, AHRQ
Rick Hermann, MD Director, Center for Quality Assessment and Improvement in Mental Health, Tufts Medical Center and UpToDate,Inc.
Neil Korsen, M.D., Medical Director, Mental Health Integration Program
Connie Horgan, Sc.D Professor and Director, Institute for Behavioral Health, Brandeis University
Harold Pincus, M.D., 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; Senior Scientist--RAND Corporation
Ben Druss M.D., M.P.H., Professor Emory University
Katherine Bradley, M.D., M.P.H Senior Investigator, Group Health Research Institute
Jeffrey Meyerhoff, M.D. National Medical Director for Medicare and Retirement, Optum Behavioral Solutions
Lisa Patton, PhD, Director of the Division of Evaluation, Analysis and Quality Center for Behavioral Health Statistics and Quality, SAMHSA
John Strauss, M.D. Medical Director Special Projects, Massachusetts Behavioral Health Partnership, A Beacon Health Options Company

Committee on Performance Measurement
Bruce Bagley, MD, American Medical Association & American Association for Physician Leadership
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Medicalis
Helen Darling, National Quality Forum
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services

David Grossman, MD, MPH, Group Health Physicians
Christine Hunter, MD (Co-Chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, United States Department of Health and Human Services (DHHS)
Nancy Lane, PhD, Vanderbilt University Medical Center
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MPH, Alliant Quality
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH Blue Shield of California
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

TECHNICAL MEASUREMENT ADVISORY PANEL
Melissa Alter, MVP Healthcare
Andy Amster, MSPH, Kaiser Permanente
Jennifer Brudnicki, MBA, Geisinger Health Plan
Kathryn Coltin, MPH, Independent Consultant
Lekisha Daniel-Robinson, Centers for Medicare and Medicaid Services
Marissa Finn, MBA, Cigna HealthCare
Scott Fox, MS, MEd, Independence Blue Cross
Carlos Hernandez, CenCal Health
Kelly Isom, MA, RN, Aetna
Harmon Jordan, ScD, RTI International
Ernest Moy, MD, MPH, Agency for Healthcare Research and Quality
Patrick Roohan, New York State Department of Health, Office of Health Insurance Programs
Lynne Rothney-Kozlak, MPH, Rothney-Kozlak Consulting, LLC
Natan Szapiro, Independence Blue Cross

HEDIS EXPERT CODING PANEL
Glen Braden, MBA, CHCA, Attest Health Care Advisors, LLC
Elonia Griffin, RN, BSN, McKesson
Denene Harper, RHIA, American Hospital Association
DeHandro Hayden, BS, American Medical Association
Patience Hoag, RHIT, CPHQ, CHCA, CCS, CCS-P, Aqurate Health Data Management
Nelly Leon-Chisen, RHIA, American Hospital Association
Tammy Marshall, LVN, Aetna
Alec McLure, RHIA, CCS-P, Verisk Health
Michele Mouradian, RN, BSN, McKesson Health Solutions
Craig Thacker, RN, CIGNA HealthCare
Mary Jane F. Toomey, RN CPC, WellCare Health Plans, Inc.

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:** 07, 2014

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, 2017

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. © 2016 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 #: 0576
Corresponding Measures:
Measure Title: Follow-Up After Hospitalization for Mental Illness (FUH)
Measure Steward: National Committee for Quality Assurance
Brief Description of Measure: The percentage of discharges for patients 6 years of age and older who were hospitalized for treatment of selected mental illness diagnoses and who had a follow-up visit with a mental health practitioner. Two rates are reported:

- The percentage of discharges for which the patient received follow-up within 30 days of discharge
- The percentage of discharges for which the patient received follow-up within 7 days of discharge.

Developer Rationale: This measure assesses whether health plan members who were hospitalized for a mental illness received a timely follow-up visit. Follow-up care following an acute event, such as hospitalization, reduces the risk of negative outcomes (e.g., medication errors, re-admission, emergency department use). Efforts to facilitate treatment following a hospital discharge also lead to less attrition in the initial post-acute period of treatment. Thus, this time period may be an important opportunity for health plans to implement strategies aimed at establishing strong relationships between patients and mental health providers and facilitate long-term engagement in treatment.

Evidence suggests that brief, low-intensity case management interventions are effective in bridging the gap between inpatient and outpatient treatment (Dixon 2009). Low-intensity interventions are typically implemented at periods of high risk for treatment dropout, such as following an emergency room or hospital discharge or the time of entry into outpatient treatment (Kreyenbuhl 2009). For example, Boyer et al evaluated strategies aimed at increasing attendance at outpatient appointments following hospital discharge. They found that the most common factor in a patient's medical history that was linked to a patient having a follow-up visit was a discussion about the discharge plan between the inpatient staff and outpatient clinicians. Other strategies they found that increased attendance at appointments included having the patient meet with outpatient staff and visit the outpatient program prior to discharge (Boyer 2000). Other studies suggest that repeated follow-up outreach and in-person visits with patients can reduce the rate of subsequent suicide attempts (Luxton, 2013) or psychiatric readmissions (Barekatain, 2014).

Barekatain M, Maracy MR, Rajabi F, Baratian H. (2014). Aftercare services for patients with severe mental disorder: A randomized controlled trial. J Res Med Sci. 19(3):240-5.

Boyer CA, McAlpine DD, Pottick KJ, Olfson M. Identifying risk factors and key strategies in linkage to outpatient psychiatric care. Am J Psychiatry. 2000;157:1592–1598.

Dixon L, Goldberg R, Iannone V, et al. Use of a critical time intervention to promote continuity of care after psychiatric inpatient hospitalization for severe mental illness. Psychiatr Serv. 2009;60:451–458.
health treatment
Kreyenbuhl, J., Nossel, I., & Dixon, L. (2009). Disengagement from mental among individuals with schizophrenia and strategies for facilitating connections to care: A review of the literature. Schizophrenia Bulletin, 35, 696-703.

Luxton DD, June JD, Comtois KA. (2013). Can postdischarge follow-up contacts prevent suicide and suicidal behavior? A review of the evidence. Crisis. 34(1):32-41. doi: 10.1027/0227-5910/a000158.

Numerator Statement: 30-Day Follow-Up: A follow-up visit with a mental health practitioner within 30 days after discharge.

7-Day Follow-Up: A follow-up visit with a mental health practitioner within 7 days after discharge.
Denominator Statement: Discharges from an acute inpatient setting (including acute care psychiatric facilities) with a principal diagnosis of mental illness during the first 11 months of the measurement year (i.e., January 1 to December 1) for patients 6 years and older.
Denominator Exclusions: Exclude from the denominator for both rates, patients who receive hospice services during the measurement year.

Exclude both the initial discharge and the readmission/direct transfer discharge if the readmission/direct transfer discharge occurs after December 1 of the measurement year.

Exclude discharges followed by readmission or direct transfer to a nonacute facility within the 30-day follow-up period regardless of principal diagnosis.

Exclude discharges followed by readmission or direct transfer to an acute facility within the 30-day follow-up period if the principal diagnosis was for non-mental health.

These discharges are excluded from the measure because rehospitalization or transfer may prevent an outpatient follow-up visit from taking place.

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

Original Endorsement Date: Dec 04, 2009 **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 *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.

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 or **Summary of prior review in [year]**

- The developer provided a rationale for timely follow-up visits after hospitalization for acute episode of mental illness.
- The developer provided NICE clinical guidelines (2009) **(guideline not graded)** related to 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 provides several guidelines related to the care and management of mental health disorders.
  - Schizophrenia: the developer provides updated NICE clinical guidelines from 2014.
    - Guideline on "preventing psychosis" and "first episode psychosis" mentions visit with mental health specialist and issues related to continuity of care.
    - Developer states the GRADE approach is used to grade the quality of evidence, and describes the process, but does not give the grade for the evidence in this guideline.
    - Developer provides process for determining strength of recommendations, but does not indicate strength of the recommendations in this guideline.
  - Schizophrenia: the developer provides APA guidelines from 2004.
    - Guideline provides several recommendations related to both the acute and stable phases of schizophrenia and the need for a combination of medication and psychosocial interventions.
    - Recommendations are recommended with either **substantial or moderate clinical confidence**.
    - While the individual studies were graded, an overall grade summary for the evidence has not been provided.
    - Processes for identifying studies, benefit, consistency, and harms are described in general, but specifics are not given for this particular guideline.
  - Bipolar Disorder: the developer provides APA guidelines from 2002.
    - "Specific goals of psychiatric management include establishing and maintaining a therapeutic alliance, monitoring the patient's psychiatric status, providing education regarding bipolar disorder, enhancing treatment compliance, promoting regular patterns of activity and of sleep, anticipating stressors, identifying new episodes early, and minimizing functional impairments **(recommended with substantial clinical confidence)**."
    - While the individual studies were graded, an overall grade summary for the evidence has not been provided.
    - Processes for identifying studies, benefit, consistency, and harms are described in general, but specifics are not given for this particular guideline.
  - Major Depressive Disorder: the developer provides APA guidelines from 2010.
    - Guideline provides several recommendations related to acute treatment and overall management of major depressive disorder; all recommendations are recommended with **substantial clinical confidence**.
    - While the individual studies were graded, an overall grade summary for the evidence has not been provided.

- Processes for identifying studies, benefit, consistency, and harms are described in general, but specifics are not given for this particular guideline.

- While all the guidelines related to the consistent and continuous management of mental illnesses, the evidence provided does not specifically address follow-up after hospitalization nor the appropriate time interval for such follow-up.
- Developer notes that (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

Exception to evidence: N/A

Questions for the Committee:

- *The evidence provided by the developer is updated and directionally the same as for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*
- *What is the relationship of this measure to patient outcomes?*
  - *How strong is the evidence for this relationship?*

Guidance from the Evidence Algorithm
Process measure based on systematic review (Box 3)→QQC not presented (Box 4)→generally strong recommendations (Box 6)→Moderate

The highest possible rating is MODERATE (due to lack of QQC provision)**.**

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.

- Performance data are summarized at the health plan level (commercial, Medicaid, and Medicare) for 2014-2016 for each of the 2 rates (7-day rate and 30-day rate) reported in this measure. Data are summarized by mean, standard deviation, minimum health plan performance, maximum health plan performance and performance at the 10th, 25th, 50th, 75th and 90th percentile. The developer also provides denominator data on the number of health plans included in HEDIS data collection and the mean eligible population for the measure across health plans.

Performance scores

Plan	Year	Mean	Standard Deviation	10th Quartile	90th Quartile	Mean Eligible Population (per plan)
7-Day Rate						
Commercial	**2014**	52.1	12.9	37.4	68.4	577
	2015	51.2	13.6	34.4	66.9	586
	2016	50.3	13.1	34.7	65.8	568
Medicaid	**2014**	34.2	14.0	17.0	54.7	928
	2015	35.2	14.7	18.0	55.8	1083
	2016	33.8	14.9	15.7	55.1	1182
Medicare	**2014**	42.0	17.0	16.5	63.2	208
	2015	43.8	16.6	20.9	63.9	245

	2016	43.6	15.7	24.7	64.2	279
30-Day Rate						
Commercial	**2014**	70.8	11.2	57.9	83.3	577
	2015	70.1	11.9	54.7	83.8	586
	2016	69.7	11.1	55.4	82.5	568
Medicaid	**2014**	60.9	18.1	32.4	80.3	935
	2015	63.0	16.1	39.4	80.2	1097
	2016	61.2	16.0	41.3	78.5	1169
Medicare	**2014**	54.2	15.0	34.6	73.1	208
	2015	54.9	15.5	36.1	76.9	245
	2016	52.4	17.0	30.6	76.2	279

Disparities

- The developer states that the CMS Office of Minority Health provides national performance data on quality measures for different racial/ethnic groups covered by Medicare in 2016. These data showed statistically significant differences in the rates for follow-up after hospitalization for a mental disorder among various racial and ethnic groups.
- The developer also provides literature about disparities in care for mental health among various in general, noting that "younger age, male gender, ethnic minority background, and low social functioning have been consistently associated with disengagement from mental health treatment.
    - The developer cites Kreyenbuhl 2009 stating that engagement strategies that specifically target high risk groups, as well as high-risk periods (including following hospital admission) can improve outcomes.

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)

1.a. Evidence to Support Measure Focus

Comments:

**a) This is an intermediate outcome of making a visit after acute hospitalization. True outcome of interest is avoiding a number of bad outcomes or continuity of care. There is good evidence linking these outcomes together.
-is there good evidence for using 7 and 30 days post hospitalization specifically?
b) Not aware of any change in evidence base. Updated evidence seems much more general in nature than is the measure, so only seems to loosely support the measure.

I agree that there's likely no need for discussion and revote of evidence for this measure.

intermediate outcomes w/ SR but w/out grading of body of evidence--> SR is likely inclusive of major evidence--> high certainty of benefit outweighing undesirable effects--> moderate

**It's not clear what type of updated literature review was done. However, I'm not aware of any research that would contradict the reliability, validity or value of this measure.

**The evidence cited supports the need for on-going treatment and stabilization of acute episodes of psychiatric illness following inpatient discharge. The desired outcome is "reduced risk of negative events – re-hospitalization, ED visits, medication errors, decline in health." I am not aware of any new information that changes the evidence base. However, I am interested into how this measure would take into account telehealth visits. The outcome being measured (appointment kept with a mental health professional within 7 and 30 days) is supported by the rationale. The literature review emphasizes the importance of case

management in allowing patients to attend their first appointment out of the hospital. The measure helps facilitate this activity so that the entity being evaluated is incentivized to provide this type of service.

**NCQA HEDIS measure % hospital discharges of patients >= 6 years with a primary dx of a mental illness plus >= 1 follow-up visit with a mh practitioner. This measure is currently in use in a number of NCQA reporting products as well as CMS programs.
MH provider spans a physician, psychologist, social worker, registered nurse, marriage and family therapist or professional counselor. The settings for a mental health visit (outpatient or partial hospitalization) include a behavioral health care facility, non-behavioral health care facility (if with mh provider or dx of mental illness) or "transitional care management services. There is no capacity to align severity of mental illness with outpatient provider type which may be clinically important given that patients who are hospitalized for a primary psychiatric diagnosis are likely to have relatively high clinical severity and complexity, and require medication treatment.

Clarification: transitional care management services are defined as?
The target population is the health plan member. The level of analysis is at the health plan level, similar to other HEDIS measures. The assumption is that the health plan is thus the point of accountability and that the health plan is responsible for an integrated health care delivery system.

However, how does this measure apply for hospitals that face the challenge of coordinating follow-up mental health care in a fragmented publicly funded mental health or outpatient mental health services that are contracted with a commercial insurer but does not have the capacity to accept all referred patients and schedule a visit within 7 or 30 days? Can poor adherence to this measure translate to any recommendations for change in the hospital discharge planning if the hospital has little authority over the resources for mental health care in the community?
In addition, NCQA acknowledges that data collection and calculation methods may vary diminishing the usefulness of HEDIS data for managed care organizations (p11, 3c, measure information). The NCQA HEDIS Compliance Audit is described, but how would this apply to providers, agencies or Medicaid MCO's that do not purchase NCQA accreditation? It appears that use by physicians for their own practices is considered a "non-commercial" use and does not require permission from NCQA to use their product. But, other users?

Further, a limitation of HEDIS measures that uses health plan level data reported to NCQA for accreditation is that there is little capacity to stratify by sociodemographic characteristics or risk adjustment. An example of an exception is work done by RAND in collaboration with CMS to examine variation in adherence to quality measures by race/ethnicity by Medicare in 2016 provided by the measure's steward. On p5 (reliability) from 2012: NCQA acknowledges, "There are no consistent standard for what entity (physician, group, plan, employer) should capture and report this data. While "requiring" reporting of the data could push the field forward, it has been our position that doing so would create substantial burden with inability to use the data because of its inconsistent." Then they restate their ability to report compliance with HEDIS measure by insurance type.

There is no evidence provided specifically for children and youth. The sources are NICE (schizophrenia adults), and APA (by target disorder: schizophrenia, bipolar, major depression. AACAP and APA recommendations are missing.

**This is a specific process measure. I have concern of the specific personal required for this measure as evidence was not presented that PCPs could not fill this role.

1a. Evidence to Support Measure Focus:
-If measuring a structure, process, or intermediate outcome: How does the evidence relate to
the specific structure, process, or intermediate outcome being measured?
There is good evidence for the measure.

-Does it apply directly or is it tangential? How does the structure, process, or intermediate outcome relate to desired outcomes?
It applies directly. If providers can get the patient to OP treatment there is evidence that readmissions are decreased.

For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission?
I am not

If measuring a health outcome or PRO: is the relationship between the measured outcome/PRO and at least one healthcare action (structure, process, intervention, or service) identified AND supported by the stated rationale?
N/A

1.b. Performance Gap

Comments:

** Data on variation among health plans we present, indicating that a measure may help systems achieve better intermediate outcomes in this area, however, no data was presented that shows that those with better follow up rates post acute hospitalization achieved lower rates of adverse outcomes or increased rates of retention in treatment. Evidence and the data provided does suggest noticeable disparity among groups by multiple factors including SES, age, social functioning and minority status.

Yes, I think there is a gap in care that warrants national measure

**Yes. There was less than optimal performance and significant variability in the measure according to HEDIS data. The only subgroups data was by Medicaid, Medicare and Commercial insurance.

**Performance data was provided and does demonstrate a gap in care. Differences are demonstrated in both the Medicare and Medicaid populations.

The developer states that the CMS Office of Minority Health provides national performance data on quality measures for different racial/ethnic groups covered by Medicare in 2016. These data showed statistically significant differences in the rates for follow-up after hospitalization for a mental disorder among various racial and ethnic groups.

**At the level of the health plan, past performance has been relatively stable for years by insurance type.
7 day
Commercial (7 days)
Ave: 2009: 54.1%; 2010: 55.96%; 2011: 57.22%; 2014 52.1%; 2015: 51.2%; 2016: 50.3%
Medicaid (7 day)
Ave: 2009: 42.6%; 2010: 42.9%; 2011: 44.6%; 2014: 34.2%; 2015: 35.2%; 2016: 33.8%
Medicare (7 day)
Ave: 2009: 37.97%; 2010: 38%; 2011: 37.8%; 2014: 42.0%, 2015: 43.8%; 2016: 43.6%
30 day
Commercial
Ave: 2009: 74.1%, 2010: 74.68%; 2011: 75.93%; 2014: 70.8%; 2015: 70.1%; 2016: 69.7%
Medicaid
Ave: 2009: 61.67%; 2010: 60.22%; 2011: 63.83%; 2014: 60.9%; 2015: 63.0%; 2016: 61.2%
Medicare
Ave: 2014: 2009: 56.32%; 2010: 55.99%; 2011: 56.69%; 54.2%; 2015: 54.9%; 2016: 52.4%

NCQA concludes findings suggest need to strength care coordination, but if no shared responsibility where do you target QI?
NCQA acknowledges that they have received questions re: research supporting this measure. (p15, 4d2.3, measure information).

**Clearly the percentage is not high and a gap which raises many other questions.

**1b. Performance Gap:
-Was performance data on the measure provided?
Yes.
-How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure?
Rates are still very low.

Disparities:
-Was data on the measure by population subgroups provided?
No as that data was not provided to the developer.
-How does it demonstrate disparities in the care?

1.c. Composite

Comments:

**Yes

**1c. Composite Performance Measure - Quality Construct (if applicable):
- Are the following stated and logical: overall quality construct, component performance measures, and their relationships; rationale and distinctive and additive value; and aggregation and weighting rules?

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Claims (only)

- The measure is specified for the health plan level of analysis and stratified by product line (Commercial, Medicare, Medicaid). The measure is based on administrative claims.
- A higher score indicates better quality.
- The measure includes 2 rates:
  - 7-Day Follow-Up: a follow up visit with a mental health practitioner within 7 days after discharge.
  - 30-Day Follow-Up: a follow-up visit with a mental health practitioner within 30 days after discharge.
- The developer provides criteria for defining a follow-up visit for the numerator.
  - The developer has provided value sets in separate Excel files.
- The denominator includes discharges from an acute inpatient setting with a principal diagnosis of mental illness during the first 11 months of the measurement year for patients 6 years and older
  - NOTE: The denominator is based on discharges, not patients.
- Denominator exclusions include:
  - Patients receiving hospice in the measurement year
  - Both the initial discharge and the readmission/direct transfer discharge if the readmission/direct transfer discharge occurs after December 1 of the measurement year.
  - Discharges followed by readmission or direct transfer to a nonacute facility within the 30-day follow-up period regardless of principal diagnosis.
  - Discharges followed by readmission or direct transfer to an acute facility within the 30-day follow-up period if the principal diagnosis was for non-mental health.
- A calculation algorithm is provided.
  - The developer provides guidance on how to identify acute inpatient discharges and readmissions to an acute inpatient care setting.

Questions for the Committee:

- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Previous measure score reliability testing (in 2012) was calculated from HEDIS data using a signal-to-noise analysis.

Describe any updates to testing:

- The developer provided a 2016 update of measure score reliability testing using a signal-to-noise analysis.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of Reliability testing

- Testing included use of a signal-to-noise analysis.
- The developer provides data on the testing sample, including the number of health plans and the mean or median eligible population per plan.

Results of reliability testing [Results of reliability testing]

Beta-binomial statistic for each measure rate

	Commercial		Medicaid		Medicare	
Year	**2012**	**2016**	**2012**	**2016**	**2012**	**2016**
7-Day Follow-Up	0.95	0.97	0.99	0.99	0.95	0.96
30-Day Follow-Up	0.97	0.96	0.99	0.99	0.95	0.97

- The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Generally, a reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities.

Questions for the Committee:

- *Does the Committee think there is a need to re-discuss and re-vote on reliability?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm [Algorithm guidance]

Specifications are precise (Box 2)→empirical reliability testing (Box 4)→score level testing (Box 5)→signal-to-noise analysis shows high signal strength (Box 6)→High

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence: The evidence is related to overall care for mental health and does not specify an optimal frequency for follow-up visit timeframes.

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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

For maintenance measures, summarize the validity testing from the prior review:

- In 2012, the developer provided results of face validity testing.
- In the previous phase, Committee members questioned the evidence supporting the 30-day timeframe and its linkage to improved outcomes. The developer noted the timeframe is supported by AAP and AACAP guidelines and that they considered other timeframes ranging from 15-45 days worked best in terms of access and claims processing issues.

Describe any updates to validity testing: No updated validity testing.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- Face validity was assessed via NCQA's standardized process called the HEDIS measure life cycle, which included field testing. Several committees of experts are engaged in this process.

Validity testing results:

Questions for the Committee:

- *No updated testing information is presented. The prior testing demonstrated good validity. Does the Committee think there is a need to re-discuss and re-vote on validity?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer does not provided data on the number of exclusions or testing.
- The developer provides guidance for identifying exclusions.
- The developer notes: "NCQA currently allows health plans for exclusion to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigate and validate the exclusion applied to the eligible denominator.

Questions for the Committee:

- o *Are the exclusions consistent with the evidence?*
- o *Are any patients or patient groups inappropriately excluded from the measure?*
- o *Have the threats to validity related to exclusions been adequately addressed?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

NCQA calculates an inter-quartile range (IQR) for each indicator, which 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 which calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The statistic is then compared against a normal distribution. If the p-value of the test statistic is less than .05, then the two plans' performances are significantly different from each other. Using this method, NCQA compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. These are updated from the 2012 rates.

ABILITY TO IDENTIFY STATISTICALLY SIGNIFICANT/MEANINGFUL DIFFERENCES

HEDIS 2016 Variation in Performance across Health Plans

Product Line	Rate	Avg. EP	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Commercial	7-Day	568	50.3%	13.1%	34.7%	42.2%	49.8%	58.7%	65.8%	16.5%	<0.001
	30-Day	568	69.7%	11.1%	55.4%	64.6%	70.6%	76.8%	82.5%	12.2%	<0.001
Medicaid	7-Day	1,182	43.6%	15.7%	24.7%	34.2%	43.6%	55.2%	64.2%	21.0%	<0.001
	30-Day	1,169	61.2%	16.0%	41.3%	54.1%	63.7%	72.6%	78.5%	18.5%	<0.001
Medicare	7-Day	279	33.8%	14.9%	15.7%	22.4%	32.0%	43.0%	55.1%	20.6%	<0.001
	30-Day	279	52.4%	17.0%	30.6%	39.8%	53.5%	65.2%	76.2%	25.4%	<0.001

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

Question for the Committee:

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

2b6. Comparability of data sources/methods:

Not needed

2b7. Missing Data

- The developer states that plans collect this measure using all administrative data sources, and asserts that NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

Guidance from the Validity Algorithm

Specifications somewhat consistent with evidence (Box 1) → potential threats to validity mostly assessed (Box2) → empirical validity testing not conducted (Box 3) → face validity systematically assessed (Box 4) → results indicate moderate agreement that the measure results can be used to distinguish quality (Box 5) →Moderate

The highest possible rating is MODERATE.

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

Committee pre-evaluation comments

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

2a.1 & 2b.1 Specifications: Reliability Specifications

Comments:

**a) what does "principal diagnosis" during first 11 months of year mean? most often listed as primary diagnosis with office visits? Doesn't use diagnosis of the admission itself? So, could be "principal diagnosis" of depression, but diagnosis for current admission could be ulcer?

b) Calculation algorithm seems clear.

c) yes, should be able to be consistently implemented

d) does not appear to be a need to revote on reliability. Yes, agree with suggested algorithm flow and suggest it to be highly reliable.

**How accurate are the data on professional type on claims? I have found them to be not completely accurate or populated.

**The specifications are clear and the measure is widely used.

**2a1. & 2b1. Specifications:
Reliability-Specifications –
-Which data elements, if any, are not clearly defined?
They are clear.

-Which codes with descriptors, if any, are not provided?
Codes seem appropriate

-Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear?
Seems logical

-What concerns do you have about the likelihood that this measure can be consistently implemented?
None

2a.2 Reliability Testing

Comments:

** yes it was tested with adequate scope for widespread implementation and with appropriate.

**Yes. Signal -to-noise ratios.

**The developer provides a detailed explanation of the reliability testing using a beta binomial approach that assesses the signal to noise ration. The results demonstrate sufficient reliability.

**Similar approach to other HEDIS measures. beta binomial method with on average score 1.0 for 7 and 30 day rates across insurance types. "Good signal to noise ratio". But is this sufficient?

**Reliable except for above comments.

**2a2. Reliability - Testing:
-Was reliability tested with an adequate scope (number of entities and patients) to generalize for widespread implementation and with an appropriate method?
Yes

-Describe how the results either do or do not demonstrate sufficient reliability.
Seemed to have good reliability testing.

-If a PRO-PM: Was testing conducted at both the data element and score levels?

-If a composite: Was testing conducted at the score level?

2b.1 Validity Specifications
Comments:
** Unclear that 7 or 30 day intervals are optimum timeframes, but sounds like it was discuss previously, so probably not a reason to re-vote.

**is there evidence that follow-up with a non-mental health professional, such as a PCP or is not effective?

**The evidence is related to overall care for mental health and does not specify an optimal frequency for follow-up visit timeframes.

**The specific of mental health professional, may cause disparities in certain populations.

**2b.1 Validity – Specifications:
-In what ways, if any, are the specifications inconsistent with the evidence?
They are consistent

-If a PRO-PM: In what ways, if any, are the specifications inconsistent with what the target population values and finds meaningful?

2b.2 Validity Testing
Comments:
** Population is large and diverse, though face validity was used including field testing and expert panel. I don't see a reason to re-discuss validity given this is unchanged from previous submission.

**Face validity.

**Face validity testing is from 2012. No updates provided.

**Face validity is based on NCQA's "measure life cycle" and "advisory panels".
The Behavioral Health panel includes only one CAP.
Meaningful differences in performance are assessed only by testing difference in adherence rates between the 25th and 75 percentile rankings. There is no comment re: the meaning of these average rates and whether this is acceptable care or related to improved clinical outcomes.

**2b2. Validity - Testing:
-Testing:
Was validity tested with an adequate scope (number of entities and patients) to generalize for widespread implementation and with an appropriate method?
Yes

Describe how the results either do or do not demonstrate sufficient validity so that conclusions about quality can be made?
Seems to be valid.

Why do you agree (or not agree) that the score from this measure as specified is an indicator of quality?
I think it provides a good way of determining which systems are functional and which are not. Can even get dowqn to the provider level i.e. those with long wait times have worse scores.

If a PRO-PM: Was testing conducted at both the data element and score levels?

2b3-7. Threats to Validity (Exclusions, Risk Adjustment, Statistically Significant Differences, Multiple Data Sources, Missing Data)
Should Note that Medicaid and dual eligible patients are not a "low touch" group.

2b3. Exclusions:
-Are the exclusions consistent with the evidence?
Yes

Are any patients or patient groups inappropriately excluded from the measure?
No

Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?
Yes

2b4. Risk Adjustment:
-If outcome (intermediate, health, or PRO-based) or resource use performance measure:
-Is there a conceptual relationship between potential SDS variables and the measure focus?
-This measure is not risk adjusted.

How well do SDS variables that were available and analyzed align with the conceptual description provided?
Developer did not provide.
-Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)?
-Was the risk adjustment (case-mix adjustment) appropriately developed and tested?
-Do analyses indicate acceptable results?
-Is an appropriate risk-adjustment strategy included in the measure?

2b5. Meaningful Differences:
-How do analyses indicate this measure identifies meaningful differences about quality?
Measure can help to identify providers that do focus on and actually have collaborative systems in place to get patients to keep their appointments.

2b6. Comparability of performance scores:
-If multiple sets of specifications:
-Do analyses indicate they produce comparable results?
-If risk-adjustment approach includes SDS factors:
-Did the developer compare performance scores with and without SDS factors in the risk-adjustment approach?
-Did the results support the risk-adjustment approach?

Not risk adjusted

2b.3.-2b7. Testing (Related to Potential Threats to Validity)
Comments:
**includes all health data and NCQA audits data for missing data, so should be good. Any exclusions to the study populations are not listed. Uninsured for lack of coverage or exclusion from policy would represent major exclusions (NCQA doe not record rationale for exclusions or who is excluded form insurance).

Agree with guidance regarding algorithm.

**not sure how reliable information on claims are regarding behavioral health professionals.

**The developer also reviews data submissions to ensure that missing data doesn't bias the entity being evaluated results.
The developer does not provided data on the number of exclusions or testing.

The developer provides guidance for identifying exclusions. Exclusions related to hospice do make sense.
The developer notes: "NCQA currently allows health plans for exclusion to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigate and validate the exclusion applied to the eligible denominator.
NCQA has a detailed methodology to identify differences among performance. NCQA calculates an inter-quartile range (IQR) for each indicator, which 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 which calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan.
Prior testing (face validity) in 2012 was determined to be good. The developer also reviews data submissions to ensure that missing data doesn't bias the entity being evaluated results.

**2b7. Missing data/no response:
-Does missing data constitute a threat to the validity of this measure?
I do not think so.

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 defined in a combination of electronic sources.
- No implementation issues were reported by the developer.
- NOTE: in the last review (2012), the Committee noted there may be difficulty following up with individuals due to socioeconomic issues (e.g., homelessness, living in group housing), which is difficult to capture in administrative data. They noted poverty, crime, and living in unsafe neighborhoods all play a role in the difficulty to ensure adequate follow-up with these patients.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are there significant circumstances that prevent follow-up in these populations?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3. Feasibility
Comments:
**Claims data should be generated if the appointment occurs. Agree w/ previous assessment that operational challenges may exist when trying to increase follow up rates (reaching folks post discharge could prove challenging based on place of living, SES, etc.)

**no concerns

**No feasibility concerns.

**see concerns under importance regarding how hospitals would identify community mental health care programs that are responsible for the follow-up care for patients discharged from their hospital, but the discharging hospital has little authority over them?

**3. Feasibility:
-Which of the required data elements are not routinely generated and used during care delivery?
-Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)?
-What are your concerns about how the data collection strategy can be put into
operational use?

No concerns

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details:

- This measure is used in the following CMS programs:
  - Medicaid Child Core Set
  - Hospital Compare
  - Electronic Health Record Incentive Program
  - CMS Physician Quality Reporting System (PQRS)
  - Physician Feedback/Quality and Resource Use Reports (QRUR)
  - Physician Value-Based Payment Modifier (VBM)
  - Inpatient Psychiatric Facility Quality Reporting Program (IPFQR)
- This measure is used by NCQA for scoring in accreditation:
  - In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries.
  - The measure is also used in the Accountable Care Organization Accreditation Program.
- This measure is also reported and used in the following:
  - State of Health Care Annual Report (nationally and by geographic region)
  - Quality Compass
  - Qualified Health Plan (QHP) Quality Rating System (QRS).
- This measure is used to calculate health plan rankings which are reported in Consumer Reports and on the NCQA website.

Improvement results:

- As shown previously, the performance rates for 2014-2016 have been generally stable across health plans, with better performance on the 30-day rate as compared to the 7-day rate.
- There is variation in performance among the health plans.

Unexpected findings (positive or negative) during implementation: None reported.

Potential harms: None reported.

Vetting of the measure:

- The developer notes they use several methods to obtain input, including several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System.
- The developer notes that health plans that report HEDIS calculate their rates and so know their performance when submitting to NCQA, and that NCQA publicly reports rates across all plans so that the plans can understand their relative performance.
- While the developers provide technical assistance for calculating/implementing the measure, It is not clear whether the developers provide specific technical assistance with interpreting the results.

Feedback:

- In 2016, the MAP Medicaid Task Force again supported the measure's continued use in the Medicaid Child Core Set.
- In 2016, the MAP Hospital Task force recommended that this measure be refined and resubmitted prior to rulemaking because it is currently specified, tested, and NQF-endorsed at the health plan level; therefore, performance on the measure cannot be attributed to the facility as currently specified. MAP agreed that a facility level measure would enable hospitals to improve follow-up care after discharge for patients with mental illness.
- The developer reported that no modifications to this measure have been required, based on feedback received.
- The developer indicated that health plans have not reported significant barriers to implementing this measure.

***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

4. Usability and Use:

Comments:

**Yes, good real world use, yes feedback has been solicited, making measurement at hospital level would be helpful as opposed to health plan, otherwise, not much has been requested in terms of change. Agree with rating of "high".

**the measure is used widely but no information or feedback from providers, patients or payers was provided to indicate that the measures is driving quality improvement.

**The measure is being publically reported including a variety of CMS programs and in the NCQA accreditation process. Feedback has been solicited. In 2016, the MAP Hospital Task force recommended that this measure be refined and resubmitted prior to rulemaking because it is currently specified, tested, and NQF-endorsed at the health plan level; therefore, performance on the measure cannot be attributed to the facility as currently specified. MAP agreed that a facility level measure would enable hospitals to improve follow-up care after discharge for patients with mental illness. The developer reported that no modifications to this measure have been required, based on feedback received. The developer indicated that health plans have not reported significant barriers to implementing this measure.

**On p5 (reliability) from 2012: NCQA acknowledges, "There are no consistent standard for what entity (physician, group, plan, employer) should capture and report this data. While "requiring" reporting of the data could push the field forward, it has been our position that doing so would create substantial burden with inability to use the data because of its inconsistent." Then they restate their ability to report compliance with HEDIS measure by insurance type.

**4. Usability and Use:
-How is the measure being publicly reported?
-For maintenance measures – which accountability applications is the measure being used for?
A number of different federal and state programs use this measure.

How can the performance results be used to further the goal of high-quality, efficient healthcare?
Could be helpful in identifying providers that need assistance in developing systems of care and developing better "access to care" strategies e.g. Same day Access models. .

Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.
I do not see any.

Has the measure been vetted in real-world settings by those being measured or others?
-If so, has data, results, and aid in interpretation been provided?
-Has feedback been solicited?
-Was feedback considered if/when changes were made to the measure?
Yes

Criterion 5: Related and Competing Measures

Related or competing measures

- 1937: Follow-Up After Hospitalization for Schizophrenia (7- and 30-day)

Harmonization

- In 2012, the Committee recommended the developer incorporate measure #1937 as a subset or target population within #0576. The developer agreed to do so following the member voting period and CSAC/Board reviews.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

The measure is not eligible for Endorsement + because empirical validity testing for the measure score has not been conducted (face validity only).

Pre-meeting public and member comments

- No comments received.

Measure Number (*if previously endorsed*)**:** 0576

Measure Title: Follow-Up After Hospitalization for Mental Illness

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: 12/2/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

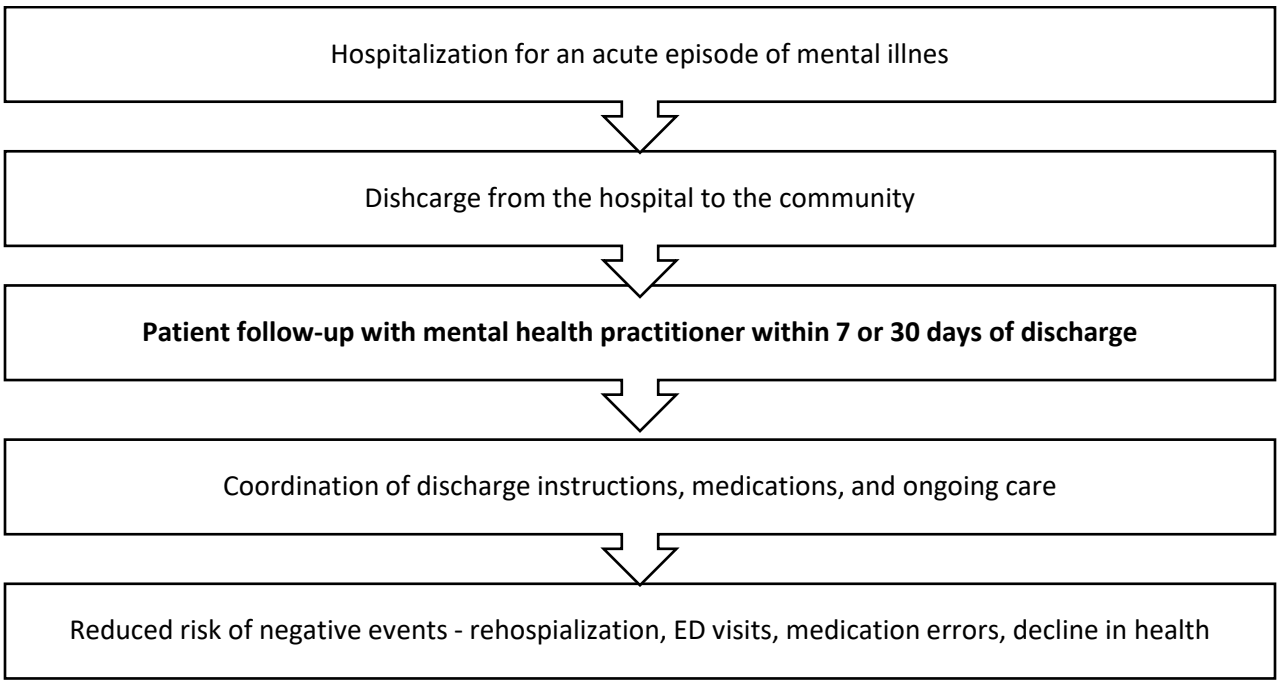
☒ Process: Follow-Up After Hospitalization for Mental Illness

☐ 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.12 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.



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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

The practice guidelines cited address treatment for the mental health diagnoses included in the measure. The guidelines provide recommendations on treatment after an acute episode of mental health disorder and recommendation on treatment to prevent (re)hospitlization.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Schizophrenia: core interventions in the treatment and management of schizophrenia in adults in primary and secondary care** National Collaborating Centre for Mental Health 2009 National Collaborating Centre for Mental Health. Schizophrenia: core interventions in the treatment and management of schizophrenia in adults in primary and secondary care. London (UK): National Institute for Health and Clinical Excellence (NICE); 2009 Mar. 41 p. (NICE clinical guideline; no. 82). http://guidelines.gov/content.aspx?id=14313
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline,	Getting Help Early • Healthcare professionals should facilitate access as soon as possible to assessment and treatment, and promote early access throughout all phases of care. Initiation of Treatment (First Episode)

summarize the conclusions from the SR.	Early Referral • Urgently refer all people with first presentation of psychotic symptoms in primary care to a local community-based secondary mental health service (for example, crisis resolution and home treatment team, early intervention service, community mental health team). Referral to early intervention services may be from primary or secondary care. The choice of team should be determined by the stage and severity of illness and the local context. • Carry out a full assessment of people with psychotic symptoms in secondary care, including an assessment by a psychiatrist. Write a care plan in collaboration with the service user as soon as possible. Send a copy to the primary healthcare professional who made the referral and the service user. • Include a crisis plan in the care plan, based on a full risk assessment. The crisis plan should define the role of primary and secondary care and identify the key clinical contacts in the event of an emergency or impending crisis. Early Post-Acute Period In the early period of recovery following an acute episode, service users and healthcare professionals will need to jointly reflect upon the acute episode and its impact, and make plans for future care.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Guideline was not graded.
Provide all other grades and definitions from the evidence grading system	N/A
Grade assigned to the **recommendation** with definition of the grade	N/A
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?	N/A
Estimates of benefit and consistency across studies	N/A
What harms were identified?	N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	N/A

National Institute for Health and Care Excellence (NICE) Guidelines

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Psychosis and schizophrenia in adults: treatment and management.** 2014 National Collaborating Centre for Mental Health. Psychosis and schizophrenia in adults: prevention and management. London (UK): National Institute for Health and Care Excellence (NICE); 2014 Mar. 58 p. (NICE clinical guideline; no 178). https://www.nice.org.uk/guidance/cg178/resources/psychosis-and-schizophrenia-in-adults-prevention-and-management-35109758952133
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.	**1.2 Preventing psychosis** 1.2.1 Referral from primary care 1.2.1.1 If a person is distressed, has a decline in social functioning and has: • transient or attenuated psychotic symptoms **or** • other experiences or behaviour suggestive of possible psychosis **or** • a first-degree relative with psychosis or schizophrenia refer them for assessment without delay to a specialist mental health service or an early intervention in psychosis service because they may be at increased risk of developing psychosis. **[new 2014]** 1.2.2 Specialist assessment • 1.2.2.1 A consultant psychiatrist or a trained specialist with experience in at-risk mental states should carry out the assessment. **[new 2014]** 1.3 First episode psychosis 1.3.1 Early intervention in psychosis services • 1.3.1.3 Early intervention in psychosis services should aim to provide a full range of pharmacological, psychological, social, occupational and educational interventions for people with psychosis, consistent with this guideline. **[2014]** • 1.3.1.4 Consider extending the availability of early intervention in psychosis services beyond 3 years if the person has not made a stable recovery from psychosis or schizophrenia. **[new 2014]** 1.3.3 Assessment and care planning • 1.3.3.1 Carry out a comprehensive multidisciplinary assessment of people with psychotic symptoms in secondary care. This should include assessment by a psychiatrist, a psychologist or a professional with expertise in the psychological treatment of people with psychosis or schizophrenia.

	1.4.6 Early post-acute period • 1.4.6.1 After each acute episode, encourage people with psychosis or schizophrenia to write an account of their illness in their notes. **[2009]** • 1.4.6.2 Healthcare professionals may consider using psychoanalytic and psychodynamic principles to help them understand the experiences of people with psychosis or schizophrenia and their interpersonal relationships. **[2009]** • 1.4.6.3 Inform the service user that there is a high risk of relapse if they stop medication in the next 1–2 years. **[2009]** • 1.4.6.4 If withdrawing antipsychotic medication, undertake gradually and monitor regularly for signs and symptoms of relapse. **[2009]** 1.4.6.5 After withdrawal from antipsychotic medication, continue monitoring for signs and symptoms of relapse for at least 2 years. **[2009]**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	For questions about the effectiveness of interventions, the GRADE approach was used to grade the quality of evidence for each outcome (Guyatt et al., 2011). For questions about the experience of care and the organisation and delivery of care, methodology checklists (see section 3.5.1) were used to assess the risk of bias, and this information was taken into account when interpreting the evidence. The technical team produced GRADE evidence profiles (see below) using GRADE profiler (GRADEpro) software (Version 3.6), following advice set out in the GRADE handbook (Schünemann et al., 2009). Those doing GRADE ratings were trained, and calibration exercises were used to improve reliability (Mustafa et al., 2013). A GRADE evidence profile was used to summarise both the quality of the evidence and the results of the evidence synthesis for each 'critical' and 'important' outcome. The GRADE approach is based on a sequential assessment of the quality of evidence, followed by judgment about the balance between desirable and undesirable effects, and subsequent decision about the strength of a recommendation. Within the GRADE approach to grading the quality of evidence, the following is used as a starting point: • RCTs 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. For the purposes of the guideline, each factor was evaluated using criteria provided in Table 4. For observational studies without any reasons for down-grading, the quality may be up-graded if there is a large

	effect, all plausible confounding would reduce the demonstrated effect (or increase the effect if no effect was observed), or there is evidence of a dose-response gradient (details would be provided under the 'other' column). Each evidence profile includes a summary of findings: number of participants included in each group, an estimate of the magnitude of the effect, and the overall quality of the evidence for each outcome. Under the GRADE approach, the overall quality for each outcome is categorised into one of four groups (high, moderate, low, very low). https://www.nice.org.uk/guidance/cg178/evidence/appendix-13-490503567
Provide all other grades and definitions from the evidence grading system	For questions about the effectiveness of interventions, the GRADE approach was used to grade the quality of evidence for each outcome (Guyatt et al., 2011). For questions about the experience of care and the organisation and delivery of care, methodology checklists (see section 3.5.1) were used to assess the risk of bias, and this information was taken into account when interpreting the evidence. The technical team produced GRADE evidence profiles (see below) using GRADE profiler (GRADEpro) software (Version 3.6), following advice set out in the GRADE handbook (Schünemann et al., 2009). Those doing GRADE ratings were trained, and calibration exercises were used to improve reliability (Mustafa et al., 2013). A GRADE evidence profile was used to summarise both the quality of the evidence and the results of the evidence synthesis for each 'critical' and 'important' outcome. The GRADE approach is based on a sequential assessment of the quality of evidence, followed by judgment about the balance between desirable and undesirable effects, and subsequent decision about the strength of a recommendation. Within the GRADE approach to grading the quality of evidence, the following is used as a starting point: • RCTs 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. For the purposes of the guideline, each factor was evaluated using criteria provided in Table 4. For observational studies without any reasons for down-grading, the quality may be up-graded if there is a large effect, all plausible confounding would reduce the demonstrated effect (or increase the effect if no effect was observed), or there is evidence of a dose-response gradient (details would be provided under the 'other' column). Each evidence profile includes a summary of findings: number of participants included in each group, an estimate of the

	magnitude of the effect, and the overall quality of the evidence for each outcome. Under the GRADE approach, the overall quality for each outcome is categorised into one of four groups (high, moderate, low, very low). https://www.nice.org.uk/guidance/cg178/evidence/appendix-13-490503567
Grade assigned to the **recommendation** with definition of the grade	The description of the process of moving from evidence to recommendations indicates that some recommendations can be made with more certainty than others. This concept of the 'strength' of a recommendation should be reflected in the consistent wording of recommendations within and across clinical guidelines. There are three levels of certainty: • recommendations for interventions that must (or must not) be used: Recommendations that an intervention must or must not be used are usually included only if there is a legal duty to apply the recommendation, for example to comply with health and safety regulations. In these instances, give a reference to supporting documents. These recommendations apply to all patients. • recommendations for interventions that should (or should not) be used: For recommendations on interventions that 'should' be used, the GDG is confident that, for the vast majority of people, the intervention (or interventions) will do more good than harm, and will be cost effective. • recommendations for interventions that could be used: For recommendations on interventions that 'could' be used, the GDG is confident that the intervention will do more good than harm for most patients, and will be cost effective Recommendations are marked as **[2009]**, **[2009, amended 2014]**, **[2014]** or **[new 2014]**. • **[2009]** indicates that the evidence has not been reviewed since 2009. • **[2009, amended 2014]** indicates that the evidence has not been reviewed since 2009 but changes have been made to the recommendation wording that change the meaning. • **[2014]** indicates that the evidence has been reviewed but no changes have been made to the recommendation. **[new 2014]** indicates that the evidence has been reviewed and the recommendation has been updated or added.

Provide all other grades and definitions from the recommendation grading system	The description of the process of moving from evidence to recommendations indicates that some recommendations can be made with more certainty than others. This concept of the 'strength' of a recommendation should be reflected in the consistent wording of recommendations within and across clinical guidelines. There are three levels of certainty: • recommendations for interventions that must (or must not) be used: Recommendations that an intervention must or must not be used are usually included only if there is a legal duty to apply the recommendation, for example to comply with health and safety regulations. In these instances, give a reference to supporting documents. These recommendations apply to all patients. • recommendations for interventions that should (or should not) be used: For recommendations on interventions that 'should' be used, the GDG is confident that, for the vast majority of people, the intervention (or interventions) will do more good than harm, and will be cost effective. • recommendations for interventions that could be used: For recommendations on interventions that 'could' be used, the GDG is confident that the intervention will do more good than harm for most patients, and will be cost effective Recommendations are marked as **[2009]**, **[2009, amended 2014]**, **[2014]** or **[new 2014]**. • **[2009]** indicates that the evidence has not been reviewed since 2009. • **[2009, amended 2014]** indicates that the evidence has not been reviewed since 2009 but changes have been made to the recommendation wording that change the meaning. • **[2014]** indicates that the evidence has been reviewed but no changes have been made to the recommendation. • **[new 2014]** indicates that the evidence has been reviewed and the recommendation has been updated or added.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	NICE guideline recommendations are based on the best available evidence. We use a wide range of different types of evidence and other information – from scientific research using a variety of methods, to testimony from practitioners and people using services.
Estimates of benefit and consistency across studies	All primary-level studies included after the first scan of citations were acquired in full and re-evaluated for eligibility at the time they were being entered into the study information

	database. More specific eligibility criteria were developed for each review question and are described in the relevant clinical evidence chapters. Eligible systematic reviews and primary-level studies were critically appraised for methodological quality (risk of bias) using a checklist (see The Guidelines Manual (NICE, 2012b) for templates). The eligibility of each study was confirmed by at least one member of the GDG.
What harms were identified?	No identified harms are cited.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

American Psychiatric Assosciation (APA) Guidelines- Schizophrenia

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition** American Psychiatric Association 2004 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
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.	Stable Phase [A, A-, B, C, D, E, F, G] "Treatment programs need to combine medications with a range of psychosocial services to reduce the need for crisis-oriented hospitalizations and emergency department visits and enable greater recovery [I]." Acute Phase Treatment [A, A-, B, C, D, E, F, G] "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 [I]." Acute Phase Treatment [A, A-, B, C, D, E, F, G] "Psychosocial interventions in the acute phase are aimed at reducing overstimulating or stressful relationships, environments, or life events and at promoting relaxation or reduced arousal through simple, clear, coherent communications and expectations; a structured and predictable environment; low performance requirements; and tolerant, nondemanding, supportive relationships with the psychiatrist and other members of the treatment team. Providing information to the patient and the family on the

	nature and management of the illness that is appropriate to the patient's capacity to assimilate information is recommended [II]. Patients can be encouraged to collaborate with the psychiatrist in selecting and adjusting the medication and other treatments provided [II]."
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	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 (see Section VI). 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
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	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 regarding the recommendation: [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?	"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?	Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

American Psychiatric Assosciation (APA) Guidelines-Bipolar Disorder

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Practice Guideline for the Treatment of Patients With Bipolar Disorder, Second Edition** American Psychiatric Association 2002 American Psychiatric Association (2002) Practice Guideline for the Treatment of Patients With Bipolar Disorder, Second Edition; 2002 Apr. 82 p. 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,	Psychiatric Management [A, C, D, E, F, G] "Specific goals of psychiatric management include establishing and maintaining a therapeutic alliance, monitoring the patient's psychiatric status, providing education regarding bipolar disorder, enhancing treatment compliance, promoting regular patterns of activity and of sleep, anticipating stressors,

summarize the conclusions from the SR.	identifying new episodes early, and minimizing functional impairments [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 (see Section VI). 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. [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	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 (see Section VI). 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
Grade assigned to the **recommendation** with definition of the grade	[I] Recommended with substantial clinical confidence.
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 regarding the recommendation: [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?	"A computerized search of the relevant literature from MEDLINE and PsycINFO was conducted. Sources of funding were not considered when reviewing the literature. The first literature search was conducted by searching MEDLINE and PsycINFO for the period from 1992 to 2000. Key words used were "bipolar disorder," "bipolar depression," "mania," "mixed states," etc. A total of 122 citations were found. A search on PubMed was also conducted through 2001 that used the search terms "electroconvulsive," "intravenous drug abuse," "treatment response," "pharmacogenetic," "attention deficit disorder," "violence," "aggression," "aggressive," "suicidal," "cognitive impairment," "sleep," "postpartum," "ethnic," "racial," "metabolism," "hyperparathyroidism," "overdose," "toxicity," "intoxication," "pregnancy," "breast-feeding," and "lactation." Additional, less formal, literature searches were conducted by APA staff and individual members of the work group on bipolar disorder"

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?	Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

American Psychiatric Assosciation (APA) Guidelines-Major Depressive Disorder

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Practice Guideline for the Treatment of Patients With Major Depressive Disorder, Third Edition** American Psychiatric Association 2010 American Psychiatric Association (2010); 2004 Practice Guideline for the Treatsment of Patients With Major Depressive Disorder, Third Edition. 2010 Oct. 151 p. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.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.	Psychiatric Management [A, A-, B, C, D, E, F, G] "Psychiatric management consists of a broad array of interventions and activities that psychiatrists should initiate and continue to provide to patients with major depressive disorder through all phases of treatment [I]." Acute Phase [A, A-, B, C, D, E, F, G] "Treatment in the acute phase should be aimed at inducing remission of the major depressive episode and achieving a full return to the patient's baseline level of functioning [I]. Acute phase treatment may include pharmacotherapy, depression-focused psychotherapy, the combination of medications and

	psychotherapy, or other somatic therapies such as electroconvulsive therapy (ECT), transcranial magnetic stimulation (TMS), or light therapy, as described in the sections that follow. Selection of an initial treatment modality should be influenced by clinical features (e.g., severity of symptoms, presence of co-occurring disorders or psychosocial stressors) as well as other factors (e.g., patient preference, prior treatment experiences) [I]. Any treatment should be integrated with psychiatric management and any other treatments being provided for other diagnoses [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 (see Section VI). 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	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 (see Section VI). 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
Grade assigned to the **recommendation** with definition of the grade	[I] Recommended with substantial clinical confidence.
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 regarding the recommendation: [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?	Relevant updates to the literature were identified through a MEDLINE literature search for articles published since the second edition of the guideline, published in 2000. For this edition of the guideline, literature was identified through a computerized search of MEDLINE, using PubMed, for the period from January 1999 to December 2006. Using the MeSH headings depression or depressive disorder, as well as the key words major depression, major depressive disorder, neurotic depression, neurotic depressive, dysthymia, dysthymic, etc. yielded 39,157 citations. An additional 8,272 citations were identified by using the key words depression or depressive in combination with the MeSH headings affective disorders or psychotic or psychosis, psychotic, catatonic, catatonia, mood disorder, etc. This yielded 13,506 abstracts, which were screened for relevance with a very modest threshold for inclusion, then reviewed by the Work Group. The Psychoanalytic Electronic Publishing database (http://www.p-e-p.org) was also searched using the terms major depression or major depressive. This search yielded 112 references. The Cochrane databases were also searched for the key word depression, and 168 meta-analyses were identified. Additional, less formal, literature searches were conducted by APA staff and individual Work Group members and included references through May 2009. Sources of funding were considered when the Work Group reviewed the literature.
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?	N/A Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this

	guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

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.

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

0576_FUH_MEF_7.0_final.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

This measure assesses whether health plan members who were hospitalized for a mental illness received a timely follow-up visit. Follow-up care following an acute event, such as hospitalization, reduces the risk of negative outcomes (e.g., medication errors, re-admission, emergency department use). Efforts to facilitate treatment following a hospital discharge also lead to less attrition in the initial post-acute period of treatment. Thus, this time period may be an important opportunity for health plans to implement strategies aimed at establishing strong relationships between patients and mental health providers and facilitate long-term engagement in treatment.

Evidence suggests that brief, low-intensity case management interventions are effective in bridging the gap between inpatient and outpatient treatment (Dixon 2009). Low-intensity interventions are typically implemented at periods of high risk for treatment dropout, such as following an emergency room or hospital discharge or the time of entry into outpatient treatment (Kreyenbuhl 2009). For example, Boyer et al evaluated strategies aimed at increasing attendance at outpatient appointments following hospital discharge. They found that the most common factor in a patient's medical history that was linked to a patient having a follow-up visit was a discussion about the discharge plan between the inpatient staff and outpatient clinicians. Other strategies they found that increased attendance at appointments included having the patient meet with outpatient staff and visit the outpatient program prior to discharge (Boyer 2000). Other studies suggest that repeated follow-up outreach and in-person visits with patients can reduce the rate of subsequent suicide attempts (Luxton, 2013) or psychiatric readmissions (Barekatain, 2014).

Barekatain M, Maracy MR, Rajabi F, Baratian H. (2014). Aftercare services for patients with severe mental disorder: A randomized controlled trial. J Res Med Sci. 19(3):240-5.

Boyer CA, McAlpine DD, Pottick KJ, Olfson M. Identifying risk factors and key strategies in linkage to outpatient psychiatric care. Am J Psychiatry. 2000;157:1592–1598.

Dixon L, Goldberg R, Iannone V, et al. Use of a critical time intervention to promote continuity of care after psychiatric inpatient hospitalization for severe mental illness. Psychiatr Serv. 2009;60:451–458.

Kreyenbuhl, J., Nossel, I., & Dixon, L. (2009). Disengagement from mental health treatment among individuals with schizophrenia and strategies for facilitating connections to care: A review of the literature. Schizophrenia Bulletin, 35, 696-703.

Luxton DD, June JD, Comtois KA. (2013). Can postdischarge follow-up contacts prevent suicide and suicidal behavior? A review of the evidence. Crisis. 34(1):32-41. doi: 10.1027/0227-5910/a000158.

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 (4b) 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 the 10th, 25th, 50th, 75th and 90th percentile. Data are stratified by year and product line (i.e. commercial, Medicaid, and Medicare). The following data demonstrate room for improvement among health plans.

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare 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.

7-Day Rate
Commercial
MEASUREMENT YEAR| N PLANS I Mean Denominator Size per plan I MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 365 I 577 I 52.1% I 12.9% I 2.8% I 88.7% I 37.4% I 45.0% I 51.9% I 60.6% I 68.4%
2015 I 355 I 586 I 51.2% I 13.6% I 3.6% I 90.9% I 34.4% I 42.3% I 51.6% I 59.6% I 66.9%
2016 I 368 I 568 I50.3% I 13.1% I 3.6% I 88.2% I 34.7% I 42.2% I 49.8% I 58.7% I 65.8%

7-Day Rate
Medicaid
MEASUREMENT YEAR| N PLANS | Mean Denominator Size per plan I MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 123 I 928 I 34.2% I 14.0% I 7.1% I 84.8% I 17.0% I 24.4% I 32.1% I 41.8% I 54.7%
2015 I 135 I 1083 I 35.2% I 14.7% I 5.7% I 84.7% I 18.0% I 24.6% I 33.0% I 43.3% I 55.8%
2016 I 166 I 1182 I 33.8% I 14.9% I 3.3% I 81.4% I 15.7% I 22.4% I 32.0% I 43.0% I 55.1%

7-Day Rate
Medicare
MEASUREMENT YEAR| N PLANS | Mean Denominator Size per plan I MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 310 I 208 I 42.0% I 17.0% I 1.6% I 76.8% I 16.5% I 31.7% I 41.9% I 54.5% I 63.2%
2015 I 300 I 245 I 43.8% I 16.6% I 1.9% I 76.3% I 20.9% I 32.0% I 45.7% I 56.8% I 63.9%
2016 I 301 I 279 I 43.6% I 15.7% I 0.0% I 75.0% I 24.7% I 34.2% I 43.6% I 55.2% I 64.2%

30-Day Rate
Commercial
MEASUREMENT YEAR| N PLANS | Mean Denominator Size per plan I MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 365 I 577 I 70.8% I 11.2% I 13.6% I 95.3% I 57.9% I 65.2% I 71.7% I 77.9% I 83.3%
2015 I 355 I 586 I 70.1% I 11.9% I 17.1% I 95.5% I 54.7% I 63.8% I 71.7% I 77.5% I 83.8%
2016 I 368 I 568 I 69.7% I 11.1% I 7.7% I 93.3% I 55.4% I 64.6% I 70.6% I 76.8% I 82.5%

30-Day Rate

Medicaid
MEASUREMENT YEAR| N PLANS | Mean Denominator Size per plan I MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 122 I 935 I 60.9% I 18.1% I 9.4% I 92.4% I 32.4% I 51.4% I 64.3% I 73.9% I 80.3%
2015 I 133 I 1097 I 63.0% I 16.1% I 7.6% I 89.2% I 39.4% I 53.2% I 66.4% I 75.1% I 80.2%
2016 I 168 I 1169 I 61.2% I 16.0% I 8.1% I 87.5% I 41.3% I 54.1% I 63.7% I 72.6% I 78.5%

30-Day Rate
Medicare
MEASUREMENT YEAR| N PLANS| Mean Denominator Size per plan I MEAN | ST DEV | MIN | MAX | 10TH | 25TH | 50TH | 75TH | 90TH
2014 I 310 I 208 I 54.2% I 15.0% I 9.3% I 89.5% I 34.6% I 43.6% I 54.7% I 64.6% I 73.1%
2015 I 300 I 245 I 54.9% I 15.5% I 8.5% I 93.7% I 36.1% I 44.3% I 53.1% I 66.7% I 76.9%
2016 I 301 I 279 I 52.4% I 17.0% I 11.1% I 88.9% I 30.6% I 39.8% I 53.5% I 65.2% I 76.2%

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 (4b) under Usability and Use.*
CMS Office of Minority Health in collaboration with the RAND Corporation provides national performance data on quality measures for different racial/ethnic groups covered by Medicare in 2016. The findings in this report indicate that Asians or Pacific Islanders that were hospitalized for mental illness more frequently had a follow-up visit with a mental health practitioner within both 7 and 30 days of discharge hospitalized for a mental disorder compared to Whites. The differences in rates for follow-up within 7 days were statistically significant ($p < 0.05$) for Asians or Pacific Islanders (40.7%) compared to Whites (35.1%). The differences in rates for and follow-up within 30 days were also statistically significant for Asian or Pacific Islanders (62.1) compared to Whites (56.3%). Blacks less frequently had a follow-up visit with a mental health practitioner after hospitalization for a mental disorder within both 7 days (25.0%) and 30 days (41.3%) of being discharged compared to Whites, 35.1% and 56.3% respectively. These differences in rates for follow-up within 7 days and 30 days were statistically significant for Blacks compared to Whites. 2014 findings indicated that Hispanics (54.3%) less frequently had a follow-up visit with a mental health practitioner within 30 days after being discharged from a hospital for a mental disorder compared to Whites (56.3%). However, the difference in rates for follow-up within 7 days of hospital discharge were not statistically significant for Hispanics (34.9%) compared to Whites (35.1%) (CMS 2016).

Centers for Medicare and Medicaid Services Office of Minority Health. (2016). Racial and Ethnic Disparities in Health Care and Medicare Advantage, Baltimore, MD.

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
Evidence from literature also shows disparities in care. Younger age, male gender, ethnic minority background, and low social functioning have been consistently associated with disengagement from mental health treatment. A recent study of disparities in follow-up after hospitalization for mental illness found that black patients were less likely than whites to receive any treatment or begin adequate follow-up within 30 days of discharge (Carson, 2014). Individuals with co-occurring psychiatric and substance use disorders, as well as those with early onset psychosis, are at particularly high risk of treatment dropout. Studies suggest that engagement strategies that specifically target these high-risk groups, as well as high-risk periods, including following an emergency room or hospital admission and the initial period of treatment, can improve outcomes (Kreyenbuhl 2009).

Carson NJ, Vesper A, Chen CN, Lê Cook B. (2014.) Quality of Follow-Up After Hospitalization for Mental Illness Among Patients From Racial-Ethnic Minority Groups. Psychiatr Serv 65(7): 888-896. doi: 10.1176/appi.ps.201300139.

Kreyenbuhl, J., Nossel, I., & Dixon, L. (2009). Disengagement from mental health treatment among individuals with schizophrenia and strategies for facilitating connections to care: A review of the literature. Schizophrenia Bulletin, 35, 696-703.

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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

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:** 0576_FUH_Value_Sets.xlsx

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 changes

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).
30-Day Follow-Up: A follow-up visit with a mental health practitioner within 30 days after discharge.

7-Day Follow-Up: A follow-up visit with a mental health practitioner within 7 days after discharge.

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).
For both indicators, a follow-up visit includes outpatient visits, intensive outpatient visits or partial hospitalizations that occur on the date of discharge. Any of the following meet criteria for a follow-up visit:
- A visit (FUH Stand Alone Visits Value Set; FUH Visits Group 1 Value Set and FUH POS Group 1 Value Set; FUH Visits Group 2 Value Set and FUH POS Group 2 Value Set) with a mental health practitioner (see definition below).
- A visit to a behavioral healthcare facility (FUH RevCodes Group 1 Value Set).
- A visit to a non-behavioral healthcare facility (FUH RevCodes Group 2 Value Set) with a mental health practitioner.
- A visit to a non-behavioral healthcare facility (FUH RevCodes Group 2 Value Set) with a diagnosis of mental illness (Mental Illness Value Set).
- Transitional care management services (TCM 7 Day Value Set).

The following meets criteria for only the 30-Day Follow-Up indicator:
- Transitional care management services (TCM 14 Day Value Set)

(See corresponding Excel document for the value sets referenced above)

Mental Health Practitioner Definition:
A practitioner who provides mental health services and meets any of the following criteria:
- An MD or doctor of osteopathy (DO) who is certified as a psychiatrist or child psychiatrist by the American Medical Specialties Board of Psychiatry and Neurology or by the American Osteopathic Board of Neurology and Psychiatry; or, if not certified, who successfully completed an accredited program of graduate medical or osteopathic education in psychiatry or child psychiatry and is licensed to practice patient care psychiatry or child psychiatry, if required by the state of practice.
- An individual who is licensed as a psychologist in his/her state of practice, if required by the state of practice.
- An individual who is certified in clinical social work by the American Board of Examiners; who is listed on the National Association of Social Worker's Clinical Register; or who has a master's degree in social work and is licensed or certified to practice as a social worker, if required by the state of practice.
- A registered nurse (RN) who is certified by the American Nurses Credentialing Center (a subsidiary of the American Nurses Association) as a psychiatric nurse or mental health clinical nurse specialist, or who has a master's degree in nursing with a specialization in psychiatric/mental health and two years of supervised clinical experience and is licensed to practice as a psychiatric or mental health nurse, if required by the state of practice.
- An individual (normally with a master's or a doctoral degree in marital and family therapy and at least two years of supervised clinical experience) who is practicing as a marital and family therapist and is licensed or a certified counselor by the state of practice, or if licensure or certification is not required by the state of practice, who is eligible for clinical membership in the American Association for Marriage and Family Therapy.
- An individual (normally with a master's or doctoral degree in counseling and at least two years of supervised clinical experience) who is practicing as a professional counselor and who is licensed or certified to do so by the state of practice, or if licensure or certification is not required by the state of practice, is a National Certified Counselor with a Specialty Certification in Clinical Mental Health Counseling from the National Board for Certified Counselors (NBCC).

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Discharges from an acute inpatient setting (including acute care psychiatric facilities) with a principal diagnosis of mental illness during the first 11 months of the measurement year (i.e., January 1 to December 1) for patients 6 years and older.

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
An acute inpatient discharge with a principal diagnosis of mental illness (Mental Illness Value Set) on or between January 1 and December 1 of the measurement year.

To identify acute inpatient discharges:
1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
2. Exclude nonacute inpatient stays (Nonacute Inpatient Stay Value Set).
3. Identify the discharge date for the stay.

The denominator for this measure is based on discharges, not on patients. If patients have more than one discharge, include all discharges on or between January 1 and December 1 of the measurement year.

Acute facility readmission or direct transfer:
If the discharge is followed by readmission or direct transfer to an acute inpatient care setting for a principal diagnosis of mental health (Mental Health Diagnosis Value Set) within the 30-day follow-up period, count only the last discharge.
To identify readmissions to an acute inpatient care setting:
1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
2. Exclude nonacute inpatient stays (Nonacute Inpatient Stay Value Set).
3. Identify the admission date for the stay.

*Due to the extensive volume of codes associated with identifying the denominator for this measure, we are attaching a separate file with value sets. See value sets located in question S.2b.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude from the denominator for both rates, patients who receive hospice services during the measurement year.

Exclude both the initial discharge and the readmission/direct transfer discharge if the readmission/direct transfer discharge occurs after December 1 of the measurement year.

Exclude discharges followed by readmission or direct transfer to a nonacute facility within the 30-day follow-up period regardless of principal diagnosis.

Exclude discharges followed by readmission or direct transfer to an acute facility within the 30-day follow-up period if the principal diagnosis was for non-mental health.

These discharges are excluded from the measure because rehospitalization or transfer may prevent an outpatient follow-up visit from taking place.

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 both the initial discharge and the readmission/direct transfer discharge if the last discharge occurs after December 1 of the measurement year.

Exclude discharges followed by readmission or direct transfer to a nonacute care setting within the 30-day follow-up period, regardless of principal diagnosis for the readmission. To identify readmissions to a nonacute inpatient care setting:
1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
2. Confirm the stay was for nonacute care based on the presence of a nonacute code (Nonacute Inpatient Stay Value Set) on the claim.
3. Identify the admission date for the stay.

Exclude discharges followed by readmission or direct transfer to an acute inpatient care setting within the 30-day follow-up period if the principal diagnosis was for non-mental health (any principal diagnosis code other than those included in the Mental Health Diagnosis Value Set). To identify readmissions to an acute inpatient care setting:
1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
2. Exclude nonacute inpatient stays (Nonacute Inpatient Stay Value Set).
3. Identify the admission date for the stay.

These discharges are excluded from the measure because rehospitalization or transfer may prevent an outpatient follow-up visit from taking place.
- See corresponding Excel document for the Value Sets referenced above in S.2b.

S.10. **Stratification Information** *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*
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 denominator. The denominator is all discharges that meet the specified denominator criteria (S7).

Step 2. Remove exclusions. Remove all discharges from the denominator that meet the specified exclusion criteria (S9).

Step 3. Identify numerator events: Search administrative systems to identify numerator events for all discharges in the denominator (S5).

Step 4. Calculate the rate by dividing the events in step 3 by the discharges in step 2.

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

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.)*
IF a PRO-PM, 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 (Only)

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 is collected.)*
IF a PRO-PM, identify the specific PROM(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 Management Organizations and Preferred Provider Organizations 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

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health : Inpatient, Behavioral Health : Outpatient, Clinician Office/Clinic
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
0576_FUH_MTF_7.0_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. (Do not remove prior testing information – include date of new information in red.)
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. (Do not remove prior testing information – include date of new information in red.)
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0576
Measure Title: Follow-Up After Hospitalization for Mental Illness
Date of Submission: 12/2/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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

1.3. What are the dates of the data used in testing? Click here to enter date range
2009-2011
2014-2016

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

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

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
2016 Update: MEASURE SCORE RELIABILITY TESTING
MEASURE SCORE RELIABILITY TESTING
The measure score reliability was calculated from 2016 HEDIS data that included 368 Commercial health plans, 166 Medicaid health plans, and 301 Medicare health plans for the 7-day follow-up rate and 368 Commercial health plans, 168 Medicaid health plans, and 301 Medicare health plans for the 30-day follow-up rate. The sample included all health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

SYSTEMATIC EVALUATION OF FACE VALIDITY
The Follow-Up After Hospitalization for Mental Illness measure was tested for face validity with several panels of experts. Measurement Advisory Panels (MAP) provide the clinical and technical knowledge required to develop the measures. The Behavioral Health MAP included 12 experts in behavioral health including representation by consumers, health plans, health care providers and policy makers. NCQA's Committee on Performance Measurement (CPM)

oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 15 members. The CPM is organized and managed by NCQA, and is responsible for advising NCQA staff on the development and maintenance of performance measures. The CPM also meets with the NCQA Board of Directors to recommend measures for inclusion in HEDIS. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement. Additional HEDIS Expert Panels provide invaluable assistance by identifying methodological issues and giving feedback on new and existing measures. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel.

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*)

2016 Update: MEASURE SCORE RELIABILITY TESTING

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

7-day Follow-Up Rate

Product Line	Number of Plans	Mean number of eligible patients per plan
Commercial	368	568
Medicaid	166	1,182
Medicare	301	279

30-Day Follow-Up Rate

Product Line	Number of Plans	Mean number of eligible patients per plan
Commercial	368	568
Medicaid	168	1,169
Medicare	301	279

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.

2016 Update: MEASURE SCORE RELIABILITY TESTING

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

Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions we have described the composition of the technical expert panel which assessed face validity in the data sample questions above.

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

2016 Update

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

2012 Submission

The measure is not stratified to detect disparities. NCQA has participated with IOM and others in attempting to include information on disparities in measure data collection. However, at the present time, this data, at all levels (claims data, paper chart review, and electronic records), is not coded in a standard manner, and is incompletely captured. There are

no consistent standards for what entity (physician, group, plan, employer) should capture and report this data. While "requiring" reporting of the data could push the field forward, it has been our position that doing so would create substantial burden with inability to use the data because of its inconsistency. At the present time, we agree with the IOM report that disparities are best considered by the use of zip code analysis which has limited applicability in most reporting situations. At the health plan level, for HEDIS health plan data collection, NCQA does have extensive data related to our use of stratification by insurance status (Medicare, Medicaid and private-commercial) and would strongly recommend this process where the data base supporting the measurement includes this information. However, we believe that the measure specifications should NOT require this since the measure is still useful where the data needed to determine disparities cannot be ascertained from the data available.

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*)

2016 Update: METHOD FOR MEASURE SCORE RELIABILITY TESTING METHOD FOR BETA-BINOMIAL RELIABILITY TESTING

The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the *signal*. The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents *noise*. The reliability of the measure is represented as the ratio of signal to noise.

- A score of 0 indicates none of the variation (signal) is attributable to the plan
- A score of 1.0 indicates all of the variation (signal) is attributable to the plan
- A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY

The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan's variation for the average plan's variation. The reliability for some plans may be more or less than the overall reliability across plans.

Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corp. TR-653-NCQA, 2009

2012 Submission

In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009) in work produced for the National Committee for Quality Assurance (NCQA).

The following is quoted from the tutorial which focused on provider-level assessment: "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." This approach is also relevant to health plans and other accountable entities.

The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. 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 accountable entity variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Adams' approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures.

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

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*)

2016 Update: MEASURE SCORE RELIABILITY

MEASURE LEVEL RELIABILITY

NCQA pools data reported by health plans according to product line. The mean reliability for the 7-day Rate per the beta binomial model was 0.97 for Commercial health plans, 0.96 for Medicare, and 0.99 for Medicaid. The mean reliability for the 30-day Rate was 0.96 for Commercial health plans, 0.97 for Medicare, and 0.99 for Medicaid.

Beta-Binomial Statisitc For Each Measure Rate

	Commercial		*Medicaid*		*Medicare*	
Rate	**Avg**	Minimum	Avg	Minimum	Avg	Minimum
7-Day Follow-Up	1.0	0.7	1.0	0.8	1.0	0.7
30-Day Follow-Up	1.0	0.6	1.0	0.8	1.0	0.8

2012 Submission

Rate 1. The percentage of members who received follow-up within 30 days of discharge
Commercial: 0.967434
Medicaid: 0.988749
Medicare: 0.949915

Rate 2. The percentage of members who received follow-up within 7 days of discharge.
Commercial: 0.954861
Medicaid: 0.989110
Medicare: 0.951935

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?*)

Among Commercial, Medicare, and Medicaid plans, results indicate both the 7-day and 30-day rates within this measure have a good signal to noise ratio that are well above the 0.7 threshold for adequate reliability. This data analysis suggests the measure has high reliability and can discriminate performance between accountable entities.

2b2. VALIDITY TESTING

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

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

☐ **Performance measure score**

☐ **Empirical validity testing**

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

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

2016 Update

Method of Assessing Face Validity

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) 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 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 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 publically 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 next year's HEDIS Volume 2.

ICD-10 CONVERSION

The below steps describe our methods to convert this measure to ICD-10 in order to develop a new code set fully consistent with the intent of the measure.

8. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use General Equivalence Mapping (GEM) to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
9. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
10. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
11. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
12. Post ICD-10 code recommendations for public review and comment.
13. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
14. NCQA staff finalize ICD-10 code recommendations.

Tools Used to Identify/Map to ICD-10

All tools used for mapping/code identification from CMS ICD-10 website (https://www.cms.gov/medicare/Coding/ICD10/index.html).
GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

Expert Participation

The NCQA HEDIS Expert Coding Panel and NCQA's Behavioral Health Measurement Advisory Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

2012 Submission

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

*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 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 MAPs, the TAG, the HEDIS Policy Panel and various other panels.

*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.Ensure funding throughout measure testing
2.Prepare a detailed conceptual and operational work-up that includes a testing proposal
3.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 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 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 Quality Compass? or in accreditation scoring.

The first-year distinction guarantees that a measure can be efficiently 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 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. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments contribute to measure evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, a third of the measurement set is researched for changes in clinical guidelines or 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 next year's HEDIS Volume 2.

What makes a measure "Desirable"?

Whether considering the value of a new measure or the continuing worth of an existing one, we must define what makes a measure useful. HEDIS measures encourage improvement. The defining question for all performance measurement—"Where can measurement make a difference?"—can be answered only after considering many factors. NCQA has established three areas of desirable characteristics for HEDIS measures, discussed below.

1. Relevance: Measures should address features that apply to purchasers or consumers, or which will stimulate internal efforts toward quality improvement. More specifically, relevance includes the following attributes.

Meaningful: What is the significance of the measure to the different groups concerned with health care? Is the measure easily interpreted? Are the results meaningful to target audiences?
Measures should be meaningful to at least one HEDIS audience (e.g., individual consumers, purchasers or health care systems). Decision makers should be able to understand a measure's clinical and economic significance.

Important to health: What is the prevalence and overall impact of the condition in the U.S. population? What significant health care aspects will the measure address?
We should consider the type of measure (e.g., outcome or process), the prevalence of medical condition addressed by the measure and the seriousness of affected health outcomes.

Financially important: What financial implications result from actions evaluated by the measure? Does the measure relate to activities with high financial impact?
Measures should relate to activities that have high financial impact.

Cost effective: What is the cost benefit of implementing the change in the health care system? Does the measure encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness?
Measures should encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness.

Strategically important: What are the policy implications? Does the measure encourage activities that use resources efficiently? Measures should encourage activities that use resources most efficiently to maximize member health.

Controllable: What impact can the organization have on the condition or disease? What impact can the organization have on the measure? Health care systems should be able to improve their performance. For outcome measures, at least one process should be controlled and have an important effect on outcome. For process measures, there should be a strong link between the process and desired outcome.

Variation across systems: Will there be variation across systems? There should be the potential for wide variation across systems.

Potential for improvement: Will organizations be able to improve performance? There should be substantial room for performance improvement.

2. Scientific soundness: Perhaps in no other industry is scientific soundness as important as in health care. Scientific soundness must be a core value of our health care system—a system that has extended and improved the lives of countless individuals.

Clinical evidence: Is there strong evidence to support the measure? Are there published guidelines for the condition? Do the guidelines discuss aspects of the measure? Does evidence document a link between clinical processes and outcomes addressed by the measure? There should be evidence documenting a link between clinical processes and outcomes.

Reproducible: Are results consistent? Measures should produce the same results when repeated in the same population and setting.

Valid: Does the measure make sense? Measures should make sense logically and clinically, and should correlate well with other measures of the same aspects of care.

Accurate: How well does the measure evaluate what is happening? Measures should precisely evaluate what is actually happening.

Risk adjustment: Is it appropriate to stratify the measure by age or another variable? Measure variables should not differ appreciably beyond the health care system's control, or variables should be known and measurable. Risk stratification or a validated model for calculating an adjusted result can be used for measures with confounding variables.

Comparability of data sources: How do different systems affect accuracy, reproducibility and validity? Accuracy, reproducibility and validity should not be affected if different systems use different data sources for a measure.

3. Feasibility:
The goal is not only to include feasible measures, but also to catalyze a process whereby relevant measures can be made feasible.

Precise specifications: Are there clear specifications for data sources and methods for data collection and reporting? Measures should have clear specifications for data sources and methods for data collection and reporting.

Reasonable cost: Does the measure impose a burden on health care systems? Measures should not impose an inappropriate burden on health care systems.

Confidentiality: Does data collection meet accepted standards of member confidentiality?
Data collection should not violate accepted standards of member confidentiality. Logistical feasibility
Are the required data available?

Auditability: Is the measure susceptible to exploitation or "gaming" that would be undetectable in an audit? Measures should not be susceptible to manipulation that would be undetectable in an audit.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
2012 Submission
Results of face validity assessment
Step 1: The Follow-Up After Hospitalization for Mental Illness measure was developed to address a gap in care concerning follow-up care for people with mental illness. NCQA's Performance Measurement Department and the Behavioral Health MAP worked together to assess the most appropriate tools for monitoring follow-up for mental illness.

Step 2: The measure was written, field-tested, and presented to the CPM and incorporated into HEDIS in 1994.

Step 3: The measure was released for Public Comment prior to publication in HEDIS. We received and responded to comments on this measure.

Step 4: The Follow-Up After Hospitalization for Mental Illness measure was introduced in HEDIS 1994. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year.

Step 5: The Follow-Up After Hospitalization for Mental Illness measure was reevaluated in 2011/2012.

2016 Update
ICD-10 CONVERSION
Summary of Stakeholder Comments Received
NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. Comments received helped to ensure we were mapping the codes correctly.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
Interpretation of systematic assessment of face validity: Our advisory panels agreed that the measures as specified will accurately differentiate quality across health plans. The measure had sufficient face validity.

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

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
2012 Submission
NCQA currently allows health plans for optional exclusion to their results. NCQA does not conduct the annual analysis applied to a sample. In measure development, field testing and any re-analysis for update, we investigate and validate the affect of the reliability exclusion applied to the eligible denominator.

2016 Update: EXCLUSIONS ANALYSIS
NCQA currently allows health plans for exclusion to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigate and validate the exclusion applied to the eligible denominator.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
N/A

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

2b4.1. What method of controlling for differences in case mix is used?
☒ No risk adjustment or stratification

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.1.1 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.

2b4.2. 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.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

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

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

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

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a*

method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)

2012 Submission

Data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance.
Comparison of means and percentiles; analysis of variance against established benchmarks: if sample size is >400, we would use an analysis of variance.

2016 Update

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 .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance using 2016 data. We used these two plans as examples of measured entities. However the method can be used for comparison of any two measured entities.

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

2016 Update

HEDIS 2016 Variation in Performance across Health Plans- Commercial

Rate	# of Plans	Avg EP	Avg.	SD	Min.	10th	25th	50th	75th	90th	IQR	P-Value
7 days	368	568	50.3%	13.1%	2.6%	34.7%	42.2%	49.8%	58.7%	65.8%	16.5%	<0.001
30 days	368	568	69.7%	11.1%	7.7%	55.4%	64.6%	70.6%	76.8%	82.5%	12.2%	<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

HEDIS 2016 Variation in Performance Across Health Plans- Medicare

Rate	# of Plans	Avg. EP	Avg.	SD	Min.	10th	25th	50th	75th	90th	IQR	P-Value
7 days	301	279	33.8%	14.9%	3.3%	15.7%	22.4%	32.0%	43.0%	55.1%	20.6%	<0.001
30 days	301	279	52.4%	17.0%	11.1%	30.6%	39.8%	53.5%	65.2%	76.2%	25.4%	<0.001

HEDIS 2016 Variation in Performan Across Health Plans- Medicaid

Rate	# of Plans	Avg. EP	Avg.	SD	Min.	10th	25th	50th	75th	90th	IQR	P-Value
7 days	166	1,182	43.6%	15.7%	0.0%	24.7%	34.2%	43.6%	55.2%	64.2%	21.0%	<0.001
30 days	168	1,169	61.2%	16.0%	8.1%	41.3%	54.1%	63.7%	72.6%	78.5%	18.5%	<0.001

Figure 1a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Commercial Plans 2014-2016

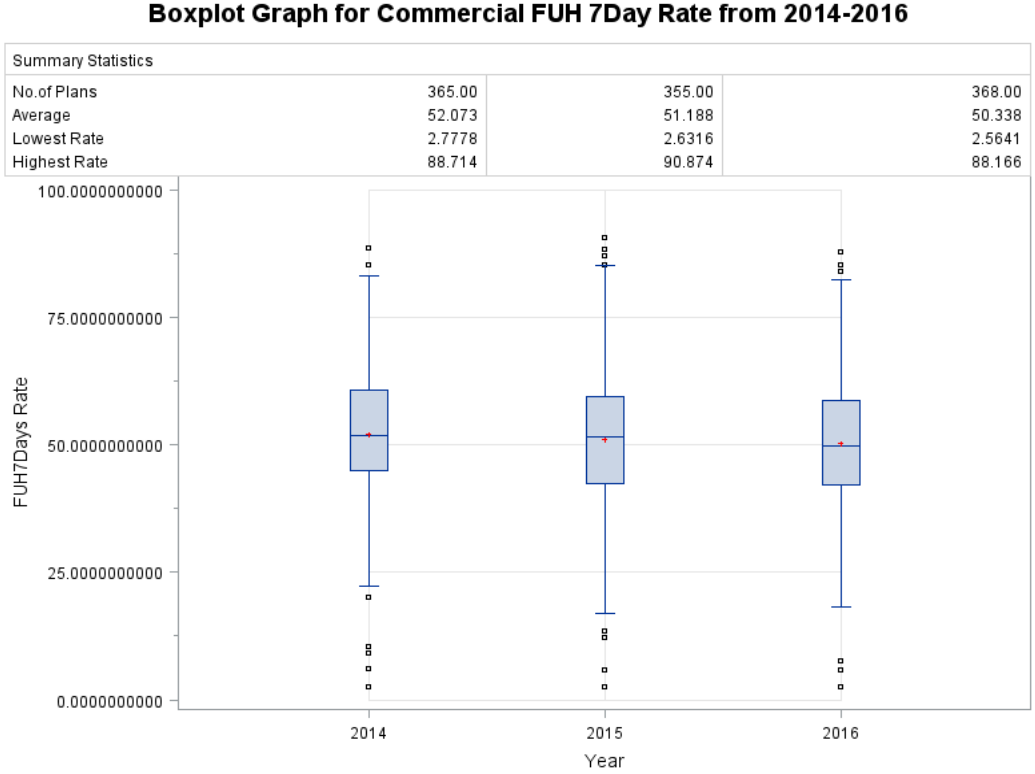


Figure 1b. *Follow-Up After Hospitalization for Mental Illness -30-Day Rate*: Commercial Plans 2014-2016

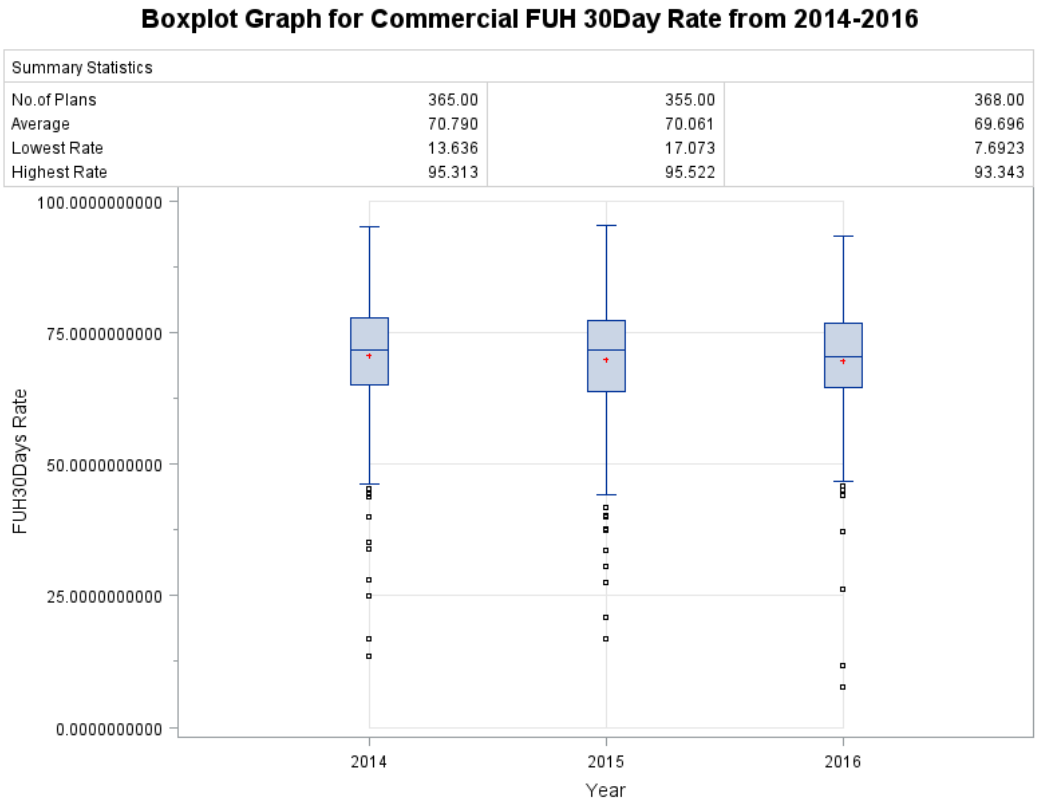


Figure 3a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Medicare Plans 2014-2016

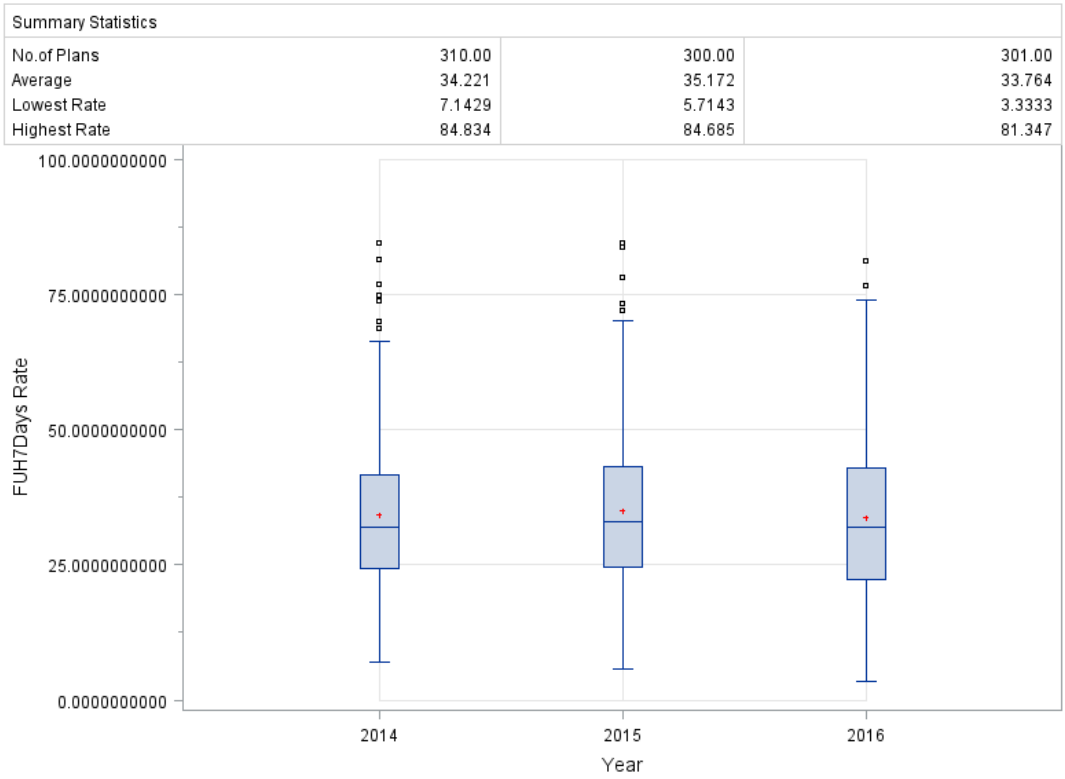


Figure 3b. *Follow-Up After Hospitalization for Mental Illness -30-Day Rate*: Medicare Plans 2014-2016

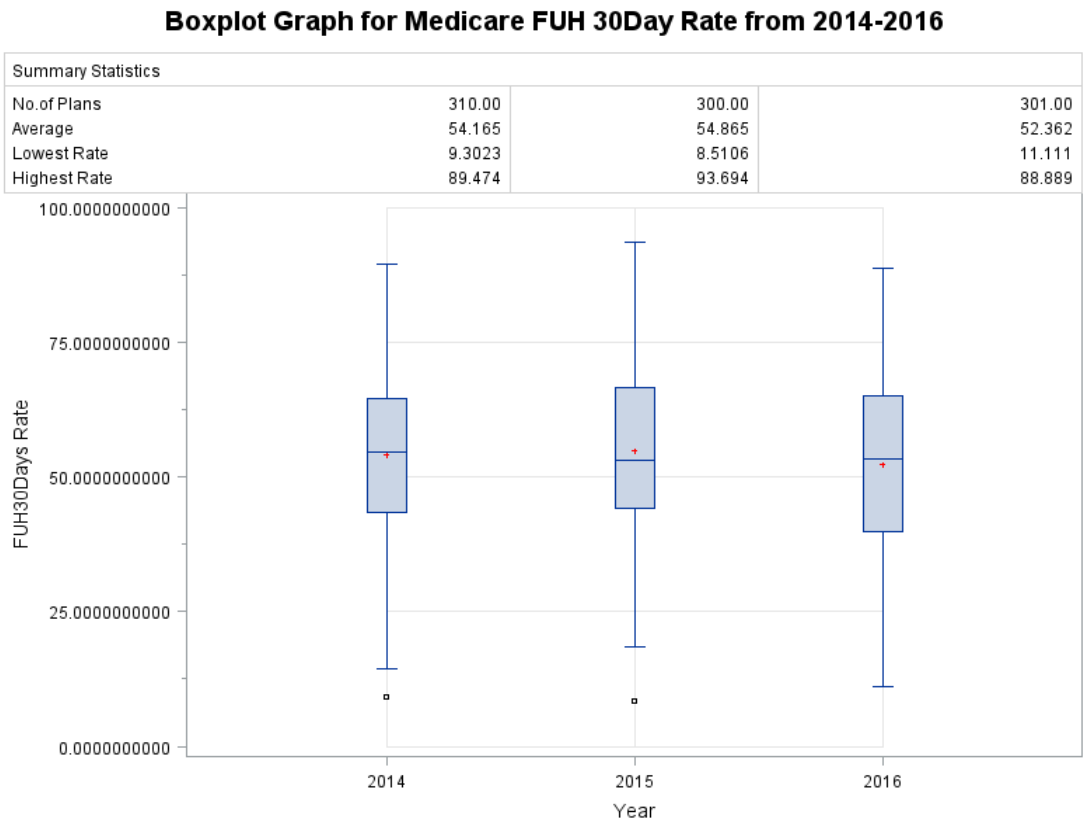


Figure 2a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Medicaid Plans 2014-2016

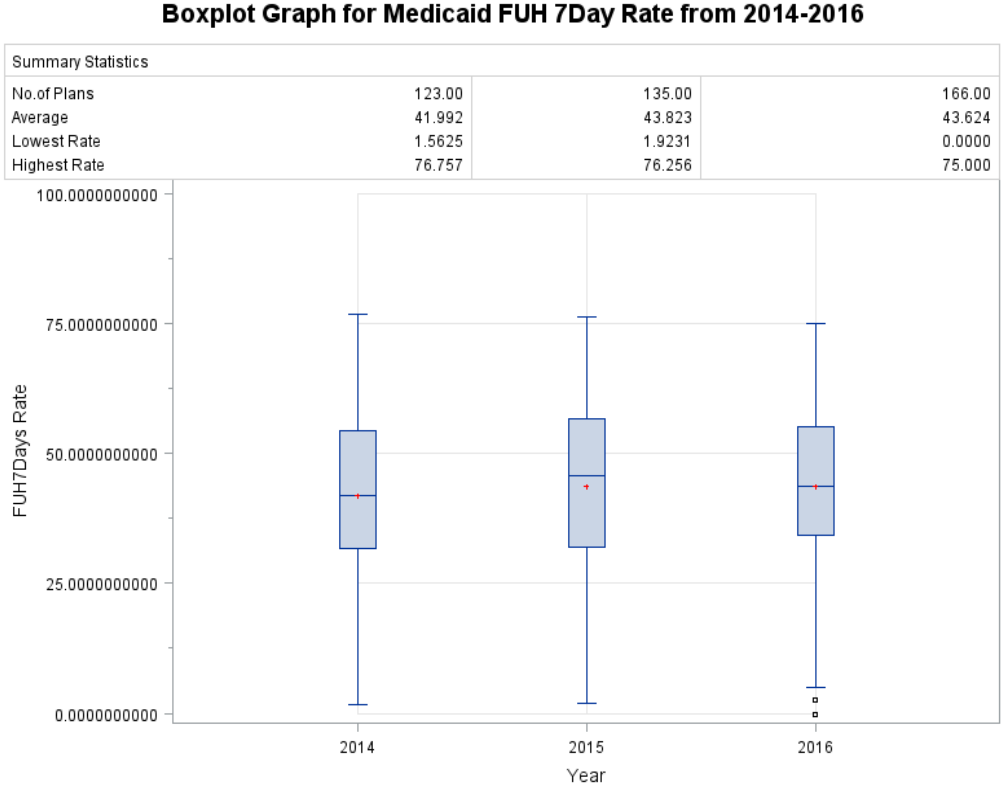
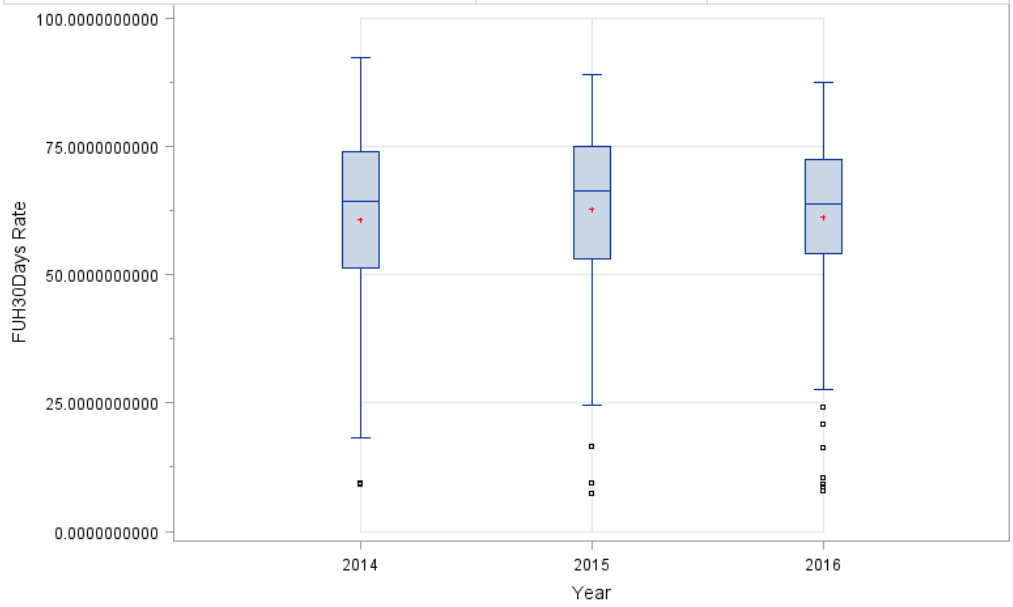


Figure 2b. *Follow-Up After Hospitalization for Mental Illness -30-Day Rate*: Medicaid Plans 2014-2016

Boxplot Graph for Medicaid FUH 30Day Rate from 2014-2016

Summary Statistics			
No.of Plans	122.00	133.00	168.00
Average	60.885	62.948	61.207
Lowest Rate	9.3750	7.6159	8.0882
Highest Rate	92.432	89.193	87.500



2012 Submission

7-Day Rate

Commercial

Measurement Year: 2009; 2010; 2011

N:	397	391	363
Min:	5.8	3.75	3.13
Max:	97.62	90.18	93.33
Mean:	54.01	55.96	57.22
SD:	13.1	13.75	12.88
P10:	37.93	39.22	42.05
P25:	45.26	46.54	48.74
P50:	53.85	56.01	57.04
P75:	62.96	65.19	66.13
P90:	71.23	72.76	72.07

Medicaid

Measurement Year: 2009; 2010; 2011

N:	62	71	85
Min:	2.6	8.2	10.87
Max:	78.57	87.9	86.85
Mean:	42.62	42.89	44.56
SD:	18.29	18.6	16.45
P10:	15.52	18.22	23.02
P25:	31.65	29.59	33.1
P50:	44.53	43.52	45.11
P75:	56.63	59.1	53.91
P90:	64.15	64.25	68.31

Medicare
Measurement Year: 2009; 2010; 2011

N:	193	231	257
Min:	4.23	2.13	1.67
Max:	86.67	84.21	84
Mean:	37.97	38	37.8
SD:	17.55	18.33	18.02
P10:	15.57	13.7	15.38
P25:	23.26	23.86	24.24
P50:	36.88	36.84	37.44
P75:	51.39	50	48.45
P90:	60.32	63.49	63.93

30-Day Rate
Commercial
Measurement Year: 2009; 2010; 2011

N:	397	391	364
Min:	21.74	21.21	13.58
Max:	98.61	97.32	100
Mean:	74.1	74.68	75.93
SD:	10.31	10.8	10.49
P10:	60	61.57	64.89
P25:	67.94	68.82	71.02
P50:	74.74	76	76.38
P75:	81.82	82.21	82.43
P90:	85.96	86.29	87.2

Medicaid
Measurement Year: 2009; 2010; 2011

N:	61	70	82
Min:	18.07	15.63	22.7
Max:	87.5	91.67	87.79
Mean:	61.67	60.22	63.83
SD:	18.25	19.14	16.19
P10:	37.27	31.79	36
P25:	49.6	49.02	57.14
P50:	64.29	62.63	66.6
P75:	75.65	74.28	74.62
P90:	81.23	83.57	82.56

Medicare
Measurement Year: 2009; 2010; 2011

N:	193	230	254
Min:	9.86	5.77	5.95
Max:	100	96.15	93.33
Mean:	56.32	55.99	56.69
SD:	18.38	19.17	18.73
P10:	30	27.3	29.79
P25:	43.82	42.11	44.87

P50:	58.1	58.23	57.95
P75:	71.43	71.88	70
P90:	78.18	79.72	80

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)
The results above indicate there is a 12-25% gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the Medicare health plans which show a 20.6-25.4 percentage point gap between 25th and 75th percentile plans.

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing*

data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each)
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased 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.**
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).

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 a PRO-PM, consider implications for both individuals providing PRO 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)
Quality Improvement (Internal to the specific organization)	Public Reporting Health Plan Ranking http://reportcard.ncqa.org/plan/external/plansearch.aspx Medicaid Child Core Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/child-core-set/index.html Medicare Adult Core Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html Hospital Compare https://www.medicare.gov/hospitalcompare/search.html? Inpatient Psychiatric Facility Quality Reporting

	http://www.qualityreportingcenter.com/wp-content/uploads/2016/06/IPF_CY2016_IPFQRManual_Guide_20160607_FINAL.pdf1_.pdf Physician Feedback/Quality and Resource Use Reports (QRUR) https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/downloads/QRUR_Presentation.pdf Qualified Health Plan (QHP) Quality Rating System (QRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/2017_QRS_and_QHP_Enrollee_Survey_Technical_Guidance.pdf Payment Program Physician Quality Reporting System https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/MeasuresCodes.html CMS EHR Incentive Program https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentiveprograms/ Physician Value-Based Payment Modifier (VBM) https://www.cms.gov/medicare/medicare-fee-for-service-payment/physicianfeedbackprogram/valuebasedpaymentmodifier.html Regulatory and Accreditation Programs Accreditation: http://www.ncqa.org/tabid/123/Default.aspx Accountable Care Organization Accreditation: http://www.ncqa.org/Programs/OtherPrograms/acomeasuresPilotProject.aspx Quality Improvement (external benchmarking to organizations) Quality Compass http://www.ncqa.org/tabid/177/Default.aspx Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

4a.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

HEALTH PLAN RANKINGS/REPORT CARDS: This measure is used to calculate health plan rakings which are reported in Consumer Reports and on the NCQA website. These rankings are based on performance on HEDIS measures among other factors. In 2012, a total of 455 Medicare Advantage health plans, 404 commercial health plans and 136 Medicaid health plans across 50 states were included in the rankings.

STATE OF HEALTH CARE ANNUAL REPORT: This measure is publically 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 2012 the report included measures on 11.5 million Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

MEDICAID CHILD CORE SET: This measure is included in the Medicaid Child Core Set which is a set of children's health care quality measures developed as part of the Children's Health Insurance Program (CHIP) Reauthorization Act for voluntary use by State

Medicaid and CHIP programs. The data collected with these measures will help CMS to better understand the quality of health care children receive through Medicaid and CHIP and assist CMS and states in moving toward a national system for quality measurement, reporting, and improvement. As per the CHIPRA legislation, state data derived from the core measures will become part of the Secretary's annual report on the quality of care for children in Medicaid and CHIP. The Secretary's annual report summarizes state-specific and national measurement information on the quality of health care furnished to children enrolled in Medicaid and CHIP.

PHYSICIAN QUALITY REPORTING SYSTEM: This measure is used in the Physician Quality Reporting System (PQRS) which is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). Eligible professionals who satisfactorily report data on quality measures for covered Physician Fee Schedule services furnished to Medicare Part B beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer) receive these payment incentives and adjustments.
CMS EHR INCENTIVE PROGRAM: This measure is used in the CMS Electronic Health Record (EHR) Incentive Program, which provides incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology.

HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. In 2012, a total of 170 Medicare Advantage health plans were accredited using this measure among others covering 7.1 million Medicare beneficiaries. [REPLACE or ADD as appropriate, 336 commercial health plans covering 87 million lives; 77 Medicaid health plans covering 9.1 million lives.] Health plans are scored based on performance compared to benchmarks.

ACCOUNTABLE CARE ORGANIZATION ACCREDITATION: This measure is used in NCQA's ACO Accreditation program, that helps health care organizations demonstrate their ability to improve quality, reduce costs and coordinate patient care. ACO standards and guidelines incorporate whole-person care coordination throughout the health care system.

QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, 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.

HOSPITAL CARE: This measure is used in Hospital Compare which helps improve quality of care by sharing objective, easy to understand data on hospital performance as well as consumer perspectives.

INPATIENT PSYCHIATRIC FACILITY QUALITY REPORTING: This measure is used in the Inpatient Psychiatric Facility Quality Reporting program which provides consumers with quality of care information to make informed decisions about their healthcare options. This program is intended to encourage clinicians and psychiatric facilities to the quality if inpatient care via awareness and reporting of best practices for respective facilities and types of care.

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.

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.

QUALIFIED HEALTH PLAN (QHP) QUALITY RATING SYSTEM (QRS): This measure is used in the Qualified Health Plan (QHP) Quality Rating System (QRS) which provides comparable information to consumers about the quality of health care services and QHP enrollee experience offered in the Marketplaces.

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

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

4b. **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 2014 to 2016, the trend across both rates and all product lines has been generally stable with some improvement in Medicaid. For the Medicaid product line the average performance has increased by two percentage points for the 7-day follow-up rate and has remained relatively stable at 62% for the 30-day follow-up rate. For the commercial product line, performance on average has been relatively stable at 51% and 70% for the 7-day and 30-day follow-up rates, respectively. For the Medicare product line, performance on average has been relatively stable at 34% for the 7-day follow-up rate and has decreased by two percentage points for the 30-day follow-up rate. Care coordination among inpatient, outpatient care settings and health plans should be strengthened in order to increase the rate of follow-up care for this vulnerable population.

4c. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **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.

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

4d1.1. **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.

4d1.2. **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, as described in Section 3c1.

4d2.1. **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.

4d2.2. **Summarize the feedback obtained from those being measured.**
In general, health plans have considered this measure feasible for reporting using the administrative data collection method. Questions received were about clarification of the specifications, such as confirmation that a type of provider met the definition of mental health providers and research supporting the measure. NCQA responded to all questions to ensure consistent implementation of the measure.

4d2.3. **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 Medicaid Child and Adult Core Sets, CMS EHR Incentive Program and CMS Physician Quality Reporting Initiative.

4d.3. **Describe how the feedback described in 4d.2 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.

5. Comparison to Related or Competing Measures

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

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

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

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

5a. Harmonization 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.**
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.
Behavioral Health Measurement Advisory Panel
Michael Schoenbaum, Ph.D., Senior Advisor for Mental Health Services, Epidemiology and Economics, National Institute of Mental Health
Frank A. Ghinassi, Ph.D., Vice President, Quality and Performance Improvement, Western Psychiatric Institute and Clinic and UPMC Behavioral Health Network, University of Pittsburgh Medical Center, Assistant Professor in Psychiatry University of Pittsburgh School of Medicine
Charlotte Mullican, B.S.W., M.P.H. Sr. Advisor for Mental Health Research, AHRQ
Rick Hermann, MD Director, Center for Quality Assessment and Improvement in Mental Health, Tufts Medical Center and UpToDate,Inc.
Neil Korsen, M.D., Medical Director, Mental Health Integration Program
Connie Horgan, Sc.D Professor and Director, Institute for Behavioral Health, Brandeis University
Harold Pincus, M.D., 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; Senior Scientist--RAND Corporation
Ben Druss M.D., M.P.H., Professor Emory University
Katherine Bradley, M.D., M.P.H Senior Investigator, Group Health Research Institute

Jeffrey Meyerhoff, M.D. National Medical Director for Medicare and Retirement, Optum Behavioral Solutions
Lisa Patton, PhD, Director of the Division of Evaluation, Analysis and Quality Center for Behavioral Health Statistics and Quality, SAMHSA
John Strauss, M.D. Medical Director Special Projects, Massachusetts Behavioral Health Partnership, A Beacon Health Options Company

Committee on Performance Measurement (CPM)
Bruce Bagley, MD, American Medical Association & American Association for Physician Leadership
Andrew Baskin, MD, Aetna
Jonathan D. Darer, MD, Medicalis
Helen Darling, National Quality Forum
Kate Goodrich, MD, MHS, Centers for Medicare and Medicaid Services
David Grossman, MD, MPH, Group Health Physicians
Christine Hunter, MD (Co-Chair), US Office of Personnel Management
Jeffrey Kelman, MMSc, MD, Centers for Medicare and Medicaid Services
Nancy Lane, PhD, Vanderbilt University Medical Center
Bernadette Loftus, MD, The Permanente Medical Group
Adrienne Mims, MD, MPH, Alliant Quality
Amanda Parsons, MD, MPH, Montefiore Health System
Eric C. Schneider, MD, MSc (Co-Chair), The Commonwealth Fund
Marcus Thygeson, MD, MPH Blue Shield of California
JoAnn Volk, MA, Georgetown University Center on Health Insurance Reforms

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 #: 0418:3132
Corresponding Measures: 0418:3148
Measure Title: Preventive Care and Screening: Screening for Depression and Follow-Up Plan
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: Percentage of patients aged 12 years and older screened for depression on the date of the encounter using an age appropriate standardized depression screening tool AND if positive, a follow-up plan is documented on the date of the positive screen
Developer Rationale: This measure aligns with the U.S. Preventive Services Task Force's (USPSTF) guidelines recommending routine screening for depression as a part of primary care for both children and adults, seeking to increase detection and treatment of depression and reduce the associated economic burden. The measure is an important contribution to the quality domain of community and population health.
The World Health Organization describes major depression as the leading cause of disability worldwide (Pratt & Brody, 2008). According to the Center for Behavioral Health Statistics and Quality (2015), in 2014 11.7 percent of adolescents aged 12 to 17 and 6.6 percent of adults 18 years and older in the United States received a diagnosis of major depressive disorder. A study by Borner et al. (2010) found that 20 percent of adolescents are likely to have experienced depression by the time they are 18 years old. In adults, depression is the leading cause of disability in high-income countries and is associated with increased mortality due to suicide and impaired ability to manage other health-related issues (Siu, 2016).
The effects of depression in adults can include difficulties in functioning at home, in the workplace, and in social situations (Pratt & Brody, 2008). For example, 35 percent of men and 22 percent of women with depression reported that their depressive symptoms make it difficult for them to work, accomplish tasks at home, or get along with other people (Pratt & Brody, 2008). Effects of depression in adolescents are similar to those in adults; however, Siu (2016) noted depression has a negative effect on developmental trajectories in children and adolescents younger than 18 years old. Also, major depressive disorder in the adolescent population is especially problematic because it is linked with higher possibility of suicide attempt, death by suicide, and recurrence of the disorder in young adulthood.
Evidence strongly recommends screening for depression in adolescent and adult patients. Specifically, the USPSTF found convincing evidence that screening improves accurate identification of adolescent and adult patients with depression in primary care settings (Siu, 2016). Yet Borner et al. (2010) cite evidence that physicians are identifying and treating depression among adolescents even less than among adults, and that more than "70 percent of children and adolescents suffering from serious mood disorders go unrecognized or inadequately treated" (Borner, 2010, p. 948). Additionally, according to the 2016 USPSTF guideline for screening for depression in children and adolescents, only 36 to 44 percent of children and adolescents with depression receive treatment, further evidence that the majority of depressed children and adolescents go untreated. Although primary care providers (PCPs) are the first line of defense in detecting depression, studies show that PCPs fail to identify up to 50 percent of depressed patients, due to both lack of time and a lack of brief, sensitive, and easy-to administer psychiatric screening tools (Borner, 2010).

Finally, according to the 2016 USPSTF guideline for screening depression among adults, the United States spent about $22.8 billion on depression treatment in 2009, and an additional estimated $23 billion on lost productivity (Siu, 2016). This substantial economic burden warrants regular screening for depression, as screening is the first step in identifying those at risk for developing major depressive disorder and closing the performance gap.

Numerator Statement: Patients screened for depression on the date of the encounter using an age appropriate standardized tool AND if positive, a follow-up plan is documented on the date of the positive screen
Denominator Statement: All patients aged 12 years and older before the beginning of the measurement period with at least one eligible encounter during the measurement period
Denominator Exclusions: Patients with an active diagnosis for Depression or a diagnosis of Bipolar Disorder are excluded.
Patients with any of the following are excepted: patient reason(s), patient refuses to participate, or medical reason(s); patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient´s health status; or situations where the patient´s functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools (for example: certain court appointed cases or cases of delirium).

Measure Type: Process
Data Source: Electronic Health Record (Only)
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.
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**

This measure is the new eMeasure version of measure 3148 (previously 0418). The information provided for Evidence is identical to that submitted for 3148. Measure 3148 will be discussed first– the ratings for evidence will automatically be assigned to this eMeasure without further discussion.

Summary of prior review of 0418 in 2014

- The claims/registry version of this eMeasure was previously evaluated as measure #0418.
- The developer cited several individual studies, literature reviews, and a consensus statement that supported use of various screening instruments, made recommendations about treatment for children and adolescents with mood disorders, provided evidence of gaps in care for post-partum women systematic reviews, and provided estimates of prevalence of depression in the U.S. They also cited a 2009 USPSTF recommendation statement for screening for depression in adults, a 2009 USPSTF systematic review on screening for depression children and adolescents in the primary care setting, a 2010 AHRQ/USPSTF clinical practice guideline recommendation for

screening of children and adolescents, and two Institute for Clinical Systems Improvement (ICSI) clinical practice guideline recommendations for screening of adults (2012) and children and adolescents (2011).

- The Committee agreed that expansion of the measure to include patients 12 to 17 years of age is supported by the USPSTF recommendation, although they noted that some primary care providers may not be able to ensure accurate diagnosis, psychotherapy, and follow-up and therefore may not screen adolescents.
- The Committee discussed, at length, the measure's requirement for annual screening for depression, noting that the USPSTF recommendation does not specify screening intervals and the other guidelines recommend different intervals. Ultimately the Committee reached consensus supporting annual screening, agreeing the potential benefits of screening outweighed the risks.
- The Committee suggested that in the future, the developer should extend the measure to consider not just whether a follow-up plan is in place, but also whether the follow-up plan is implemented.

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:

- In their logic model, the developers link screening for depression to receipt of follow-up care, which they then link to improved health and quality of life.
- 2016 USPSTF recommendation statement on screening for depression in children and adolescents (**moderate quality evidence, "B" recommendation**).
- 2016 USPSTF recommendation statement on screening for depression in adults (**moderate quality evidence, "B" recommendation**).
- 2016 ICSI guideline recommendations for treating depression in adults in the primary care setting (**low quality evidence, strong recommendation for low quality evidence**)

Questions for the Committee:

- *The evidence for this measure is identical to #3148. Is there a reason to vote again?*

Guidance from the Evidence Algorithm

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

The highest possible rating is 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 provided annual average performance rates for 2011-2014, based on data from the PQRS program, and CY 2015 distributional statistics from EHR data from a convenience sample of two practices.

PQRS data, 2011-2014 (all reporting methods)

Year	Average performance	Percent of eligible professionals reporting
2011	82.6%	0.6%

2012	65.2%	0.4%
2013	71.0%	1.3%
2014	52.4%	7.5%

Performance Rates, EHR Data (convenience sample), Calendar Year 2015 data

Source	# providers	# patients	Mean	Standard deviation	10th percentile	30th percentile	Median	70th percentile
EHR	57	54,349	68.8%	20.2%	40.9%	67.5%	72.6%	79.4%

Disparities

- The developer provided patient-level data from the convenience sample. These results indicate that prevalence of screening varies according to age group, race, and sex, with lower rates in younger patients, blacks and whites, and males.
- The developer also cited recent literature indicating lower rates of screening and treatment in minority adults and lower rates of screening among men.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist [for screening for depression screening and documenting a follow-up plan] in other patient subpopulations?*

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 to Support Measure Focus

Comments:

**Agree with good evidence that depression should be screened, assessed, and treated accordingly. there should be a better method for training of those in primary care.
**Other than the frequency issue, I see no reason this e-measure should not be as linked causally to improved outcomes as the parent measure.
**The evidence for this measure is covered in measure #3148.
**Measuring the process (in this case, screening for depression) leads to early identification and treatment, which is critical to prevent re-occurrence and suicide.

1b. Performance Gap

Comments:

**No additional comments.
**Yes.
**Yes, a gap in care exists among gender, age, and race. Data provided on different age groups, as well as racial and ethnic groups.
**There is considerable variation and less then optimal performance in screening; disparities exist. There is great opportunity for improvement.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Electronic health record

Specifications:

- This measure is specified for the individual clinician and clinical group/practice levels of analysis in the clinician office/clinic setting. A higher score indicates better quality.
- The numerator includes the number of patients screened using an age-appropriate standardized tool and, if positive, those for whom a follow-up plan is documented at the time of the positive screen.
    - A listing of examples of screening tools that would be appropriate for meeting the measure is provided.
    - Documentation of a follow-up plan must include at least one of the following: additional evaluation for depression; Suicide Risk Assessment; referral to a practitioner who is qualified to diagnose and treat depression; pharmacological interventions; or other interventions or follow-up for the diagnosis or treatment of depression.
- The denominator includes patients age 12 or older with a clinician office/clinic encounter.
- Exclusions include encounters where: the patient refuses to participate; the patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status; situations where the patient's functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools; patient has an active diagnosis of depression; or patient has a diagnosed bipolar disorder.
- A straightforward calculation algorithm is provided, which should allow for consistent calculation of the measure.

Questions for the Committee:

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

eMeasure Technical Advisor(s) review (if not an eMeasure, delete this section):

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	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; OR
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

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Reliability of the measure score was assessed using EHR data from two practices in Pennsylvania (one primary care and one pediatrics) for encounters from 1/1/2015 to 12/31/2015.

Practice	Specialty	EHR Vendor	# of Providers	# Patients
A	Primary care	GE Centricity 12.0	53	52,961
B	Pediatrics	Medent 22.0	4	1,388

- Score-level testing using a signal-to-noise analysis (using the beta-binomial method) was conducted.
  - This type of analysis, which is an appropriate method for demonstrating score-level reliability, quantifies the amount of variation in performance that is due to differences between providers compared to differences due to measurement error. Results will vary based on the amount of variation between the providers and the number of patients treated by each provider. This method results in a reliability statistic that ranges from 0 to 1. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

- NOTE: For testing, clinicians with <10 reporting events in the measurement period were excluded from the analysis. However, the measure specifications do not limit the measure to those with at least 10 reporting events. Therefore, reliability estimates from the analysis likely are higher than would be found if all clinicians were included (as typically, reliability increases with increased sample size).

Results of reliability testing

Data source	Number of providers	Between-provider variance	Reliability mean	Reliability median	Reliability Std dev	Reliability min/max
EHR	52	.028	0.984	0.995	0.045	0.724 – 1.000

Note: Four providers were dropped from the reliability analysis who had 20 or fewer eligible patients.

Questions for the Committee:

- *Are the test samples adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing with measure as specified (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → High certainty that measure results are reliable (Box 6a) → High

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence The evidence does not specify an optimal frequency for screening.

Question for the Committee:

- *Given a lack of evidence regarding an optimal frequency for screening for depression, do you agree that screening at each visit is reasonable?*

2b2. Validity testing

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

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**
- ☒ **Bonnie Testing**

Validity testing method:

- Bonnie testing
  - At this time, NQF accepts Bonnie testing in lieu of empirical validity testing for "legacy" eMeasures (measures that have been respecified into eMeasures and are currently used in federal quality programs).
  - The developer created a test deck of 22 synthetic patient records were used to test the measure logic and value sets. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure.
- Face-validity testing
  - Developers asked 12 clinicians who were eligible to report on the measure to rate their agreement regarding whether the measure results will provide an accurate reflection of quality and can be used to distinguish good and poor quality (scale 1-5).
  - The clinicians included in the face validity assessment were not involved in the development of the measure.

Validity testing results:

- Bonnie testing

- Results reached 100% coverage and confirmed there was a test case for each pathway of logic (negative and positive test cases).
- The measure also had a 100% passing rate which confirmed that all the test cases performed as expected.

- Face-validity testing
  - Nine of the 12 clinicians (75%) agreed or strongly agreed that the measure accurately reflects care quality, while 3 (25%) disagreed. Those who disagreed noted issues related to patient compliance, burden of documentation, and desire for use of only one specified tool for screening for adolescents.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Exclusions include encounters where: the patient refuses to participate; the patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status; situations where the patient's functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools; patient has an active diagnosis of depression; or patient has a diagnosed bipolar disorder.
- The developer tested the frequency of exclusions using EHR data extracts from the two testing sites.
  - The rate of exclusions and exceptions was 12.1% for all patients reported by the two practices.
  - The vast majority were patients identified as having been previously diagnosed with depression.
  - The developer states this is consistent with research on lifetime prevalence for depression and justifies the use of the exclusions/exceptions to account for situations in which it is appropriate not to screen and follow-up with patients for depression.

Questions for the Committee:

- *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)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

- The developer used EHR data from the two practices to calculate measure performance scores and assess the distribution of performance.
- The provider notes that these results represent only those providers who participated in the testing of this measure may not be generalizable to the population of all eligible providers.

Performance Rates, EHR Data (convenience sample), Calendar Year 2015 data

# providers	# patients	Mean	Standard deviation	10th percentile	30th percentile	Median	70th percentile
57	54,349	68.3%	20.2%	40.9%	67.5%	72.6%	79.4%

Average performance score, by practice (1/1/2015-12/31/2015)

Practice	Number of providers	Average weighted score	Average unweighted score
A	53	68.2%	70.7%
B	4	71.1%	70.5%

Question for the Committee:

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

2b6. Comparability of data sources/methods:

- Not needed, but the developer notes that they designed the specifications to maximize alignment and consistency with measure 3225 (based on data from claims/registry).

2b7. Missing Data

- The developer provides an explanation of the method used to identify missing data.
- The developer notes that both practices had issues with reporting of specific interventions (often captured in "additional evaluation" field) and optional variables (exceptions).
- The developer notes that of the 54,349 patient records:
  - 14 (<1%) did not have a reported sex
  - 67 (<1%) did not have a reported provider
- The developer reports that they did include patients whose race and/or ethnicity were unknown in the analysis of disparities; results were comparable to those of patient not Hispanic or Latino.

Guidance from the Validity Algorithm

Specifications are mostly consistent with the evidence (frequency not addressed) (Box 1) → Threats to validity assessed → Bonnie testing performed in lieu of empirical validity testing, as allowed (Box 3) → Face validity testing (Box4) and empirical testing of data elements using Bonnie tool (Box 10) →Method appropriate for legacy eMeasures (Box 11) →Moderate

The highest possible rating is MODERATE (because score-level testing was not conducted).

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

Comments:

**Reliability appears good; no additional comments.
**Yes, although one wonders if varied EHRs code and abstract such information the same--just idle curiosity...
**Data elements are clearly defined. The calculation algorithm is clear. It is likely the measure can be consistently implemented. Test samples are adequate to generalize widespread implementation.
**The measure is well defined and specified. The measure can be consistently implemented.

2a2. Reliability Testing

Comments:

**No additional comments
**OK
**There were 52 providers used in testing reliability with a high level of reliability.
**Reliability was assessed from two practices (one primary care and one pediatrics); good reliability.

2b1. Validity Specifications

Comments:

**appears valid; no additional comments.
**Yes, this seems as valid as has been discussed, other than the frequency.

**Depending on the tool used, screening at each visit is reasonable.
**There is consistency with the evidence.

2b2. Validity Testing
Comments:
**No additional comments
**OK
**The sample size for testing face validity (12) seems low.
**The measure has face validity; also Bonnie testing. Measure score validity testing has not been conducted. Validity-moderate.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**I question if patients choose not to fill out screening tool, how does this get accounted for?
**Not big issues, although SDS probably important impacts.
**No, missing data is not a threat to validity.
**Rate of exclusion was 12.1% with the majority being previously diagnosed with depression. Risk adjustment was none.

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 required data elements are routinely collected and are available in electronic sources. The developer tested the measure in three EHRs and was found to be mostly feasible, as the elements required are in structured data fields.
- There was some concern about identifying follow-up interventions, or those in the denominator exceptions (e.g., patient refused), but they concluded that these elements are unlikely to be used frequently enough to compromise the feasibility of the measure.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
**agree with comment about it may be difficult (depending on how charting is done) to identify follow up specificity of effectiveness of follow up recommendations for those with positive screens.
**Yes, feasible
**Data elements are available in electronic form.
**Collection of data through EMRs is feasible although identification of follow-up could be challenging.

Criterion 4: Usability and Use

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- This measure is used the CMS Physician Quality Reporting System (PQRS) *[which is being phased out by 12/31/18 and is replaced by MIPS]*.
  - In 2014, 7.5% of >500,000 eligible professionals reported on the measure. However, The developer does not state how many of these report using the eMeasure option
- The measure also is included in the following CMS programs:
  - Electronic Health Record Incentive Program: EPs
  - Medicare Shared Savings Program (MSSP)
  - Merit-Based Incentive Payment System (MIPS) Program
  - Physician Value-Based Payment Modifier (VBM) *[which is being phased out by 12/31/18 and is replaced by MIPS]*
  - Physician Feedback/Quality and Resource Use Reports (QRUR) *[which is being phased out by 12/31/18 and is replaced by MIPS]*
  - Physician Compare
  - Medicaid Adult Core Set

Improvement results

PQRS data, 2011-2014 (indicates data reported through EHRs, claims, and registries). Performance rates based only on the eMeasure have not been provided.

Year	Average Performance	Percent of eligible professionals reporting
2011	82.6%	0.6%
2012	65.2%	0.4%
2013	71.0%	1.3%
2014	52.4%	7.5%

Unexpected findings (positive or negative) during implementation None reported, other than those described when conducting workflow assessments.

Potential harms: None reported.

Vetting of the measure: None reported.

Feedback:

Feedback based on the claims/registry version (#3148):

- In 2012, the Measure Applications Partnership (MAP) supported the measure for inclusion in the clinician Meaningful Use program and also specifically referenced the measure as relevant to the dual eligible beneficiary population.
- In 2012, HHS included in measure in the CMS initial core set of measures for Medicaid-eligible adults.
- In 2014, the MAP supported the measure for inclusion in the Physician Compare and Value-Based Payment Modifier programs. The MAP also supported its inclusion in the End-Stage Renal Disease Quality Incentive Program, even though the measure is specified for individual clinicians/practices and not for dialysis facilities. The MAP Medicaid Task Force also supported its continued use in the Medicaid Adult Core Set.
- In 2015, the MAP Medicaid Task Force again supported the measure's continued use in the Medicaid Adult Core Set.

***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?*
- o *Has the measure been vetted in real-world settings by those being measure or others?*
- o *Does inclusion of the measure in the QRUR program meet the requirements for vetting by those being measured?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

**I would like to see evidence of screening improving outcomes. I would like to see improved training for those entering primary care. (it's one thing to screen; it's another thing to feel comfortable in further assessment and knowing who to refer to and when to refer).
**Yes, I suspect.
**Measure is publicly reported using PQRS. It is unclear how many providers are reporting using eMeasure. The measure is also reported through various CMS programs.
**PQRS; Number of eligible professional reports has increased from .6% in 2011 to 7.5% in 2014, but still a large number of professionals are not reporting the measure.

Criterion 5: Related and Competing Measures

Competing measures:

0518: Depression Assessment Conducted

3148: Preventative Care and Screening: Screening for Depression and Follow-Up Plan

- #3148 is the claims-based version of this measure. NQF will not ask the Committee to select a best-in-class measure.

Harmonization:

- Measure #0518 is a measure that is specified for facility-level assessment in the home health setting and thus the Committee will not be asked to select a best-in-class measure.
  - During the 2014 evaluation of the measure, the developers agreed that there were opportunities for harmonization, including adding the home health setting to this measure (#3132) and adding a follow-up requirement to #0518. The Committee recommended that the developers pursue harmonizing the measures in the areas of care settings and follow-up.
  - No change to the care setting for this measure has been made.
- It appears that the claims-based "version"(#3148) has been harmonized with #3132 to the extent possible and thus NQF will not ask the Committee to discuss harmonization.

Endorsement + Designation
The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.
This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users. **Eligible for Endorsement + designation**: ☐ **Yes** ☒ **No** **RATIONALE IF NOT ELIGIBLE**: The measure is not eligible for Endorsement + because measure score validity testing has not been conducted.

Pre-meeting public and member comments

- None received.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0418 (3132/3148)

Measure Title: Preventive Care and Screening: Screening for Depression and Follow-Up Plan

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: None

Date of Submission: 12/9/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

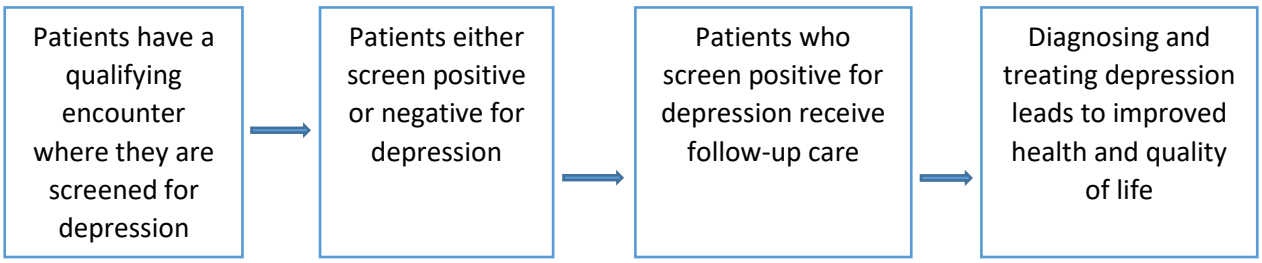
☒ Process: Depression screening and follow-up plan

☐ 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.12 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.



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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

N/A

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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**	• **Screening for Depression in Children and Adolescents: U.S. Preventive Services Task Force Recommendation Statement** • **Albert L. Siu, MD, MSPH, on behalf of the U.S. Preventive Services Task Force** • **Published February 9, 2016** • **U.S. Preventive Services Task Force. Screening for depression in children and adolescents: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2016 Mar 1;164(5):360-366** • **https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/depression-in-children-and-adolescents-screening1?ds=1&s=depression**
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 USPSTF recommends screening for major depressive disorder (MDD) in adolescents aged 12 to 18 years. Screening should be implemented with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up. (B recommendation)" (p. 360)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Moderate: "The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies; inconsistency of findings across individual studies; limited generalizability of findings to routine primary care practice; and lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion." (p. 367)
Provide all other grades and definitions from the evidence grading system	As indicated in Appendix Table 2 (p. 360). High: "The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies." Low: "The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: the limited number or size of studies; important flaws in study design or methods; inconsistency of findings across individual studies; gaps in the chain of evidence; findings that are not generalizable to routine primary care practice; and a lack of information on important health outcomes. More information may allow an estimation of effects on health outcomes."

Grade assigned to the **recommendation** with definition of the grade	B Recommendation: "The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial." (p. 367)
Provide all other grades and definitions from the recommendation grading system	As indicated in Appendix Table 1 (p. 360). A: "The USPSTF recommends the service. There is high certainty that the net benefit is substantial." C: "The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small." D: "The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits." I statement: "The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined."
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The USPSTF conducted a systematic evidence review to update its 2009 recommendation on screening for child and adolescent major depressive disorder (MDD) in primary care settings. Compared to its 2009 review, USPSTF narrowed the scope of this evidence review to focus exclusively on screening for and treating MDD. The USPSTF excluded studies of paroxetine because in 2003 the U.S. Food and Drug Administration (FDA) recommended not to use paroxetine to treat MDD in children and adolescents. The USPSTF examined evidence on the benefits and harms of screening; the accuracy of screening tests feasible for use in primary care; and the benefits and harms of treatment with psychotherapy, medications, and collaborative care models in patients ages 7 to 18 years. USPSTF limited treatment studies to those that were implemented in or required referrals from primary care settings to ensure that the patient population was similar to patients who would be identified through screening. The USPSTF found five good- or fair-quality studies of the accuracy of MDD screening instruments in children and adolescents ages 11 years or older. It also found eight good- or fair-quality randomized controlled trials (RCTs) that reported health outcomes in children or adolescents with screen-detected MDD: four for patients who were treated with selective serotonin reuptake inhibitors (SSRIs), two involving psychotherapy, one on SSRIs combined with psychotherapy, and one on collaborative care. Most trials were restricted to adolescents ages 12 to 14 years or older; only two of the fourSSRI trials included children ages 7 or 8 years (p. 364).
Estimates of benefit and consistency across studies	The USPSTF found adequate evidence that screening tests can accurately identify MDD in adolescents and that treatment of adolescents with screen-detected MDD is associated with beneficial reductions in symptoms. The USPSTF therefore concluded with moderate certainty that screening for MDD in adolescents ages 12 to 18 years is associated a moderate net benefit (p. 365). Although the USPSTF found no studies that directly evaluated whether screening for MDD in adolescents in primary care settings leads to improved health and other outcomes, there is adequate evidence that treatment of MDD detected through screening in adolescents is associated with moderate benefit, for example, by reducing the severity of depression or improving depression symptoms (p. 361).

What harms were identified?	The USPSTF found no direct evidence to suggest that screening or treatment for MDD in adolescents or children leads to potential harms. Seven trials in the USPSTF review pertaining to the use of SSRIs (five trials), psychotherapy with or without SSRIs (one trial), or collaborative care (one trial) reported on harms, but none found significant differences between intervention groups. The USPSTF noted, however, that some of the studies had insufficient power to detect differences on some of the measured outcomes.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	This guideline was published in 2016 and is the most recent systematic review completed.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Screening for Depression in Adults US Preventive Services Task Force Recommendation Statement** • **Albert L. Siu, MD, MSPH; and the US Preventive Services Task Force (USPSTF)** • **Published January 26, 2016** • **US Preventive Services Task Force (USPSTF). Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. JAMA. 2016; 315(4):380-387. doi:10.1001/jama.2015.18392.** • **https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/depression-in-adults-screening1?ds=1&s=depression**
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 USPSTF recommends screening for depression in the general adult population, including pregnant and postpartum women. Screening should be implemented with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up (B recommendation)" (p. 360).
Grade assigned to the **evidence** associated with the recommendation	As indicated in Appendix Table 2 (p. 367). Moderate: "The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies; inconsistency of findings across individual studies; limited generalizability of

with the definition of the grade	findings to routine primary care practice; and lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion."
Provide all other grades and definitions from the evidence grading system	As indicated in Appendix Table 2 (p. 367). High: "The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies." Low: "The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: the limited number or size of studies; important flaws in study design or methods; inconsistency of findings across individual studies; gaps in the chain of evidence; findings that are not generalizable to routine primary care practice; and a lack of information on important health outcomes. More information may allow an estimation of effects on health outcomes."
Grade assigned to the **recommendation** with definition of the grade	As indicated in Appendix Table 1 (p. 367). Graded B Recommendation: "The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial."
Provide all other grades and definitions from the recommendation grading system	As indicated in Appendix Table 1 (p. 367). A: "The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Offer or provide this service." C: "The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small." D: "The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits." I statement: "The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined."
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The USPSTF found convincing evidence that screening tests accurately identify depression in the general adult population, and that there are benefits to treating depression once diagnosed. Nine good- or fair-quality trials addressed screening in general adults and older adults (p.384). The evidence from five RCTs, in addition to indirect evidence reviewed for the 2009 recommendation, indicates that there is moderate certainty that screening for depression in adults is of moderate benefit (p. 385). In terms of treatment, two systematic reviews concluded that antidepressants were effective in treating depression in older adults. Two good-quality systematic reviews found that older adults who received psychotherapy were more than twice as likely to have remission as those who received no treatment (p. 384). For pregnant and postpartum women, the USPSTF reviewed 23 studies comparing the accuracy of the Edinburgh Postnatal Depression Scale with diagnostic interview, and found that the instrument had an acceptable positive predictive

	value for detecting MDD. The USPSTF identified six good- or fair-quality RCTs that assessed the effect of screening for depression in pregnant and postpartum women that support recommending depression screening for this group (p. 384). Eighteen trials examined the benefits of treatment interventions in women who screened positive for depression in primary care or community settings. Of these, 15 focused on postpartum women and 3 involved pregnant women. Ten RCTs found that cognitive behavioral therapy (CBT) benefits both postpartum and pregnant women; the remaining eight trials did not find sufficient evidence to draw conclusions about the effectiveness of treatment in these populations (p. 385). The USPSTF review found seven studies that compared suicide-related events in adults who received SSRIs and other antidepressants versus placebo. None of the studies showed a significant increase in suicide completion among adults taking antidepressants, but given the rarity of this event, they may not have had sufficient power to detect differences between treatment groups (p. 385). The majority of the evidence on the harms of antidepressants in pregnant and postpartum women comes from a good-quality comprehensive systematic review on the comparative effectiveness and safety of antidepressant treatment for depression in this population. The review, which included 124 observational studies, showed that second-generation antidepressant use during pregnancy may be associated with a small increase in the risk of several outcomes, including preeclampsia, miscarriage, and respiratory distress (p. 385).
Estimates of benefit and consistency across studies	The USPSTF concluded with at least moderate certainty that there is a moderate net benefit to screening for depression in adults, including older adults, and in pregnant and postpartum women who receive care in clinical practices that have CBT or other evidence-based counseling available after screening (p. 381). The USPSTF also found adequate evidence that programs that screen for depression and have adequate support systems in place improve clinical outcomes in both the general adult population and among pregnant and postpartum women specifically. Improvement in outcomes included reduction and remission of depression symptoms (p. 380).
What harms were identified?	“The USPSTF found adequate evidence that the magnitude of harms of screening for depression in adults is small to none. The USPSTF found adequate evidence that the magnitude of harms of treatment with CBT in postpartum and pregnant women is small to none” (p. 380). “The USPSTF found that second-generation antidepressants (mostly selective serotonin reuptake inhibitors [SSRIs]) are associated with some harms, such as an increase in suicidal behaviors in adults aged 18 to 29 years and an increased risk of upper gastrointestinal bleeding in adults older than 70 years, with risk increasing with age; however, the magnitude of these risks is, on average, small. The USPSTF found evidence of potential serious fetal harms from pharmacologic treatment of depression in pregnant women, but the likelihood of these serious harms is low. Therefore, the USPSTF concludes that the overall magnitude of harms is small to moderate” (p. 381).
Identify any new studies conducted	This guideline was published in 2016 and is the most recent systematic review completed.

since the SR. Do the new studies change the conclusions from the SR?	

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Health Care Guideline Adult Depression in Primary Care** • **Institute for Clinical Systems Improvement (ICSI)** • **March 2016** • **Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, Mack N, Myszkowski M. Institute for Clinical Systems Improvement. Adult Depression in Primary Care. Updated March 2016. Pp.1-131** • **https://www.icsi.org/guidelines__more/catalog_guidelines_and_more/catalog_guidelines/catalog_behavioral_health_guidelines/depression/**
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.	All recommendations are contained within a table on pages 8-10. We have included below those that are most relevant to this measure: "Clinicians should routinely screen all adults for depression using a standardized instrument" (p. 8). "Clinicians should establish and maintain follow-up with patients" (p. 10). "Clinicians should screen and monitor depression in pregnant and post-partum women" (p. 10).
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	All three of the recommendations listed above were graded as Low Quality Evidence: "Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain." (p. 4)
Provide all other grades and definitions from the evidence grading system	ICSI utilizes the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology system. Visit http://www.gradeworkinggroup.org/ for more information about GRADE. High Quality Evidence: **"**Further research is very unlikely to change our confidence in the estimate of effect." (p. 4) Moderate Quality Evidence: "Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate." (p. 4)
Grade assigned to the	Strong Recommendation for Low Quality Evidence: "The work group feels that the evidence consistently indicates the benefit of this action outweighs the harms. This

recommendation with definition of the grade	recommendation might change when higher quality evidence becomes available." (p. 4)
Provide all other grades and definitions from the recommendation grading system	Strong Recommendation for High Quality Evidence: "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." (p. 4) Weak Recommendation for High Quality Evidence: "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." (p. 4) Strong Recommendation for Moderate Quality Evidence: "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." (p. 4) Weak Recommendation for Moderate Quality Evidence: "The work group recognizes that there is a balance between harms and benefits, 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." (p. 4) Weak Recommendation for Low Quality Evidence: "The work group recognizes that there is significant uncertainty about the best estimates of benefits and harms." (p. 4)
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The authors used a consistent and defined literature search process to develop and revise the ICSI guidelines. First, ICSI staff, in consultation with the work group and a medical librarian, conducted a literature search to identify systematic reviews, randomized clinical trials, meta-analyses, other guidelines, regulatory statements, and any other pertinent literature. Work group members then evaluated the identified literature using the GRADE methodology (p. 131). For this guideline, ICSI reviewed 12 systematic reviews, 17 meta-analyses, 18 RCTs, 1 meta-regression, and 2 guidelines. The body of evidence related to screening all adults and pregnant and post-partum women was of low to moderate quality. The body of evidence for establishing a follow-up plan was of high quality.
Estimates of benefit and consistency across studies	For the recommendation: "Clinicians should routinely screen all adults for depression using a standardized instrument," ICSI determined that screening results in finding and treating more depressed patients, leading to better outcomes and improved functioning not only for depression, but also for other co-morbid conditions. There is also some evidence that screening may reduce overall, long-term medical costs for depressed patients (p. 14). For the recommendation: "Clinicians should establish and maintain follow-up with patients," ICSI determined that appropriate, reliable follow-up is highly correlated with improved response and remission scores. Follow-up is also correlated with the improved safety and efficacy of medications and helps prevent relapse (p. 50).

	For the recommendation: "Clinicians should screen and monitor depression in pregnant and post-partum women," ICSI determined that screening patients leads clinicians to find and treat more patients with depression. Furthermore, untreated prenatal depression is associated with negative pregnancy outcomes such as poor maternal self-care, poor nutrition, preterm labor, and low birth weight. Untreated prenatal depression is also associated with negative effects on children such as developmental delay and cognitive impairment (p. 122).
What harms were identified?	The only harms identified for all three recommendations were the cost of screening patients who are not depressed, and the potential additional cost of unnecessary follow-up visits (p. 14, p. 50, p. 122).
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We have not identified any additional new studies published since the release of this guideline that would change its conclusions.

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.

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_form_NQF_0418-636178223236625911.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

This measure aligns with the U.S. Preventive Services Task Force's (USPSTF) guidelines recommending routine screening for depression as a part of primary care for both children and adults, seeking to increase detection and treatment of depression and reduce the associated economic burden. The measure is an important contribution to the quality domain of community and population health.

The World Health Organization describes major depression as the leading cause of disability worldwide (Pratt & Brody, 2008). According to the Center for Behavioral Health Statistics and Quality (2015), in 2014 11.7 percent of adolescents aged 12 to 17 and 6.6 percent of adults 18 years and older in the United States received a diagnosis of major depressive disorder. A study by Borner et al. (2010) found that 20 percent of adolescents are likely to have experienced depression by the time they are 18 years old. In adults, depression is the leading cause of disability in high-income countries and is associated with increased mortality due to suicide and impaired ability to manage other health-related issues (Siu, 2016).

The effects of depression in adults can include difficulties in functioning at home, in the workplace, and in social situations (Pratt & Brody, 2008). For example, 35 percent of men and 22 percent of women with depression reported that their depressive symptoms make it difficult for them to work, accomplish tasks at home, or get along with other people (Pratt & Brody, 2008). Effects of depression in adolescents are similar to those in adults; however, Siu (2016) noted depression has a negative effect on developmental trajectories in children and adolescents younger than 18 years old. Also, major depressive disorder in the adolescent population is especially problematic because it is linked with higher possibility of suicide attempt, death by suicide, and recurrence of the disorder in young adulthood.

Evidence strongly recommends screening for depression in adolescent and adult patients. Specifically, the USPSTF found convincing evidence that screening improves accurate identification of adolescent and adult patients with depression in primary care settings (Siu, 2016). Yet Borner et al. (2010) cite evidence that physicians are identifying and treating depression among adolescents even less than among adults, and that more than "70 percent of children and adolescents suffering from serious mood disorders go unrecognized or inadequately treated" (Borner, 2010, p. 948). Additionally, according to the 2016 USPSTF guideline for screening for depression in children and adolescents, only 36 to 44 percent of children and adolescents with depression receive treatment, further evidence that the majority of depressed children and adolescents go untreated. Although primary care providers (PCPs) are the first line of defense in detecting depression, studies show that PCPs fail to identify up to 50

percent of depressed patients, due to both lack of time and a lack of brief, sensitive, and easy-to administer psychiatric screening tools (Borner, 2010).
Finally, according to the 2016 USPSTF guideline for screening depression among adults, the United States spent about $22.8 billion on depression treatment in 2009, and an additional estimated $23 billion on lost productivity (Siu, 2016). This substantial economic burden warrants regular screening for depression, as screening is the first step in identifying those at risk for developing major depressive disorder and closing the performance gap.

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 (4b) under Usability and Use.*
Provider-level performance scores suggest that there are still gaps in care and opportunities for improvement.
Average Performance Rates by Year (PQRS – all reporting methods)*:
2011–82.6% (0.6% of eligible professionals reporting)
2012–65.2% (0.4% of eligible professionals reporting)
2013–71.0% (1.3% of eligible professionals reporting)
2014–52.4% (7.5% of eligible professionals reporting)
*From the 2014 PQRS Reporting Experience Report and Appendix

EHR data from convenience sample of two practices, 1/1/2015 through 12/31/2015
Number of Providers 57
Number of patients 54,349
Average Unweighted Score 70.7%
Average Weighted Score 68.3%

Standard deviation 20.2%
Minimum 0.0%
Interquartile range 18.2%
10th percentile 40.9%
20th percentile 61.5%
30th percentile 67.5%
40th percentile 70.0%
Median 72.6%
60th percentile 77.0%
70th percentile 79.4%
80th percentile 86.6%
90th percentile 93.8%
Maximum 100.0%

Please note: The unweighted average measure is the aggregated score for entire population. The weighted average is the average provider-level score, which is weighted by the number of patients in the denominator of each provider's score. All other statistics are based on weighted provider-level scores.

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 (4b) under Usability and Use.

Below are aggregate performance rates by patients' age, race, ethnicity, and sex from a convenience sample of two practices' EHR data from 2015. Because practices were not able to provide data on patients' insurance status, socioeconomic status, and/or disability status, we were unable to determine the presence of disparities in performance based on these factors. These results represent only those providers who participated in the testing of this measure and may not be generalizable to the population of all eligible providers.

Age Groups
12–17: 53.7%
18–64: 58.3%
65+: 91.4%
(X2 = 5,252.569; df = 2; N = 47,782; p < 0.0001)

Race
American Indian or Alaska Native: 46.2%
Asian: 52.9%
Black: 72.4%
Native Hawaiian or other Pacific Islander: 51.4%
White: 69.4%
Multiracial: 72.2%
Unknown: 58.7%
(X2 = 270.069; df = 6; N = 47,782; p < 0.0001)

Ethnicity
Hispanic or Latino: 59.6%
Not Hispanic or Latino: 68.4%
Unknown: 66.5%
(X2 = 15.823; df = 2; N = 47,782; p = 0.0004)

Sex
Female: 68.5%
Male: 68.0%
(X2 = 1.362; df = 1; N = 47,768; p = 0.2431)
We excluded 14 patients whose sex was unknown

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

Race/ethnicity: Literature indicates that depression rates are higher in non-Latino black people than in their non-Latino white counterparts (Pratt & Brody, 2008). Clinical practice guidelines also indicate that minority racial and cultural groups in the United States are less likely to receive treatment for depression than European Americans (Trangle et al., 2016). Data collected from electronic health records of approximately 65,079 adult primary care patients from 2010 to 2012 showed that (1) individuals from minority groups are less likely to undergo screening for mental disorders, such as depression screening; (2) minority groups have less access to mental health care and receive less than adequate health care compared to non-Latino whites, and (3) women from racial/ethnic minority groups are less likely than white women to have access to mental health care (Hahm et al., 2015). Medicare beneficiary survey data analyzed by Akincigil et al. showed that about 6.4 percent of whites, 4.2 percent of black Americans, and 7.2 percent of Latino Americans had a diagnosis of depression. Among those diagnosed, 73 percent of whites received treatment (either with antidepressants, psychotherapy, or both); 60 percent of black Americans received treatment; and 63.4 percent of Latino Americans received treatment (Akincigil et al., 2012). These findings are consistent with other studies that show depression is under-recognized and undertreated among adult minorities. According to Davis et al. (2011), "Recent data suggest that the proportion of depressed adults who seek treatment is significantly lower among African Americans (53%) than among Caucasians (67%)."

Age: Literature indicates that depression rates are highest among adults ages 40 to 59 (Pratt & Brody, 2008).
Gender: Literature indicates that depression is more common in women than in men (Pratt & Brody, 2008). Studies showed that men were less likely than women to receive screening for mental health problems, such as depression (Hahm et al., 2015). Among Latino and Asian Americans, women were more likely than men to receive screening for depression and visit a health care provider for depression care after depression was detected. Asian and black Americans, particularly black women, were less likely to receive screening for depression and less likely to receive any depression care (Hahm et al., 2015).
Socioeconomic status: People with incomes below the federal poverty line and in the 18-39 and 40-59 age brackets experience higher depression rates than those with higher incomes, although this disparity is not observable in other age categories (Pratt & Brody, 2008).

We did not find any literature related to disparities associated with insurance status or disability.

Akincigil, A., Olfson, M., Siegel, M., Zurlo, K. A., Walkup, J. T., & Crystal, S. (2012). Racial and ethnic disparities in depression care in community-dwelling elderly in the United States. American Journal of Public Health, 102, 2, 319-328.
Davis, T. D., Deen, T., Bryant-Bedell, K., Tate, V., & Fortney, J. (2011). Does minority racial-ethnic status moderate outcomes of collaborative care for depression? Psychiatric Services, 62, 1282-1288.
Hahm, H. C., Cook, B. L., Ault-Brutus, A., & Alegri´a, M. (2015). Intersection of race-ethnicity and gender in depression care: Screening, access, and minimally adequate treatment. Psychiatric Services, 66, 258-264.
Pratt, L. A., & Brody, D. J. (2008). Depression in the United States household population, 2005-2006 (NCHS Data Brief No. 7). Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
Trangle, M., Gursky, J., Haight, R., Hardwig, J., Hinnenkamp, T., Kessler, D., Myszkowski, M. (2016, March). Adult depression in primary care. Bloomington, MN: Institute for Clinical Systems Improvement. Retrieved from https://www.icsi.org/guidelines__more/catalog_guidelines_and_more/catalog_guidelines/catalog_behavioral_health_guidelines/depression/

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, 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.)*
https://ecqi.healthit.gov/ep/ecqms-2017-performance-period/preventive-care-and-screening-screening-depression-and-follow-plan

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:** CMS2v6.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:** NQF_0418_Coding_Table_S2b._CMS_2.xlsx

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.
Since the last annual NQF measure update in 2015, we made minor changes to the measure specifications. Changes for program year 2016 include: added Patient Health Questionnaire (PHQ-9) and Pediatric Symptom Checklist (PSC-17) to the Definition section of the specification; added examples of depression screening tools to clarify available standardized options for provider use, including depression screening tools for adolescents; changed term clinical depression to depression because the word clinical could reduce the sensitivity of screening; incorporated new literature into rationale. We also completed several coding updates, including: updated QDM data elements and logic to conform to standards; deleted one CPT code (90839) from the list of eligible encounters; added one RXNORM code (259197) and deleted ten RXNORM codes (107078, 242345, 242637, 242715, 252718, 259993, 309671, 309672, 410062, 991200) from list of depression medications for adults; added one ICD9CM code (296.25) to list of bipolar diagnoses.

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 screened for depression on the date of the encounter using an age appropriate standardized tool AND if positive, a follow-up plan is documented on the date of the positive screen

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).
Within the eMeasure specification, value sets contain various codes to indicate clinical quality actions. (See attached code table for S2.b)
Definitions included in relation to the numerator include the following:
Screening – Completion of a clinical or diagnostic tool used to identify people at risk of developing or having a certain disease or condition, even in the absence of symptoms.
Standardized Depression Screening Tool – A normalized and validated depression screening tool developed for the patient population in which it is being utilized. Examples of adolescent depression screening tools (12 – 17 years) include but are not limited to: Patient Health Questionnaire for Adolescents (PHQ-A), Beck Depression Inventory-Primary Care Version (BDI-PC), Mood Feeling Questionnaire (MFQ), Center for Epidemiologic Studies Depression Scale (CES-D), Patient Health Questionnaire (PHQ-9), Pediatric Symptom Checklist (PSC-17), PRIME MD-PHQ2.

Examples of adult depression screening tools (18 years and older) include but are not limited to Patient Health Questionnaire (PHQ9), Beck Depression Inventory (BDI or BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Depression Scale (DEPS), Duke Anxiety-Depression Scale (DADS), Geriatric Depression Scale (GDS), Cornell Scale Screening, PRIME MD-PHQ2.
Follow-Up Plan - Documented follow-up for a positive depression screening must include one or more of the following:
•Additional evaluation for depression
•Suicide Risk Assessment
•Referral to a practitioner who is qualified to diagnose and treat depression

•Pharmacological interventions
•Other interventions or follow-up for the diagnosis or treatment of depression

The measure specification defines the numerator as:
AND:
- OR:
  - AND: Most Recent: "Occurrence A of Risk Category Assessment: Adolescent Depression Screening (result)" during ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")
  - AND: "Occurrence A of Risk Category Assessment: Adolescent Depression Screening (result: Negative Depression Screening)"
  - AND: Age< 18 year(s) at: "Measurement Period"
- OR:
  - AND: Most Recent: "Occurrence A of Risk Category Assessment: Adolescent Depression Screening (result)" during ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")
  - AND: "Occurrence A of Risk Category Assessment: Adolescent Depression Screening (result: Positive Depression Screening)"
  - AND: Union of:
    "Intervention, Performed: Additional evaluation for depression - adolescent"
    "Intervention, Order: Referral for Depression Adolescent"
    "Medication, Order: Depression medications - adolescent"
    "Intervention, Performed: Follow-up for depression - adolescent"
    "Procedure, Performed: Suicide Risk Assessment"
    <= 1 day(s) starts after or concurrent with start of "Occurrence A of Risk Category Assessment: Adolescent Depression Screening"
  - AND: Age< 18 year(s) at: "Measurement Period"
- OR:
  - AND: Most Recent: "Occurrence A of Risk Category Assessment: Adult Depression Screening (result)" during ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")
  - AND: "Occurrence A of Risk Category Assessment: Adult Depression Screening (result: Negative Depression Screening)"
  - AND: Age>= 18 year(s) at: "Measurement Period"
- OR:
  - AND: Most Recent: "Occurrence A of Risk Category Assessment: Adult Depression Screening (result)" during ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")
  - AND: "Occurrence A of Risk Category Assessment: Adult Depression Screening (result: Positive Depression Screening)"
  - AND: Union of:
    "Intervention, Performed: Additional evaluation for depression - adult"
    "Intervention, Order: Referral for Depression Adult"
    "Medication, Order: Depression medications - adult"
    "Intervention, Performed: Follow-up for depression - adult"
    "Procedure, Performed: Suicide Risk Assessment"
    <= 1 day(s) starts after or concurrent with start of "Occurrence A of Risk Category Assessment: Adult Depression Screening"

AND: Age>= 18 year(s) at: "Measurement Period"

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 12 years and older before the beginning of the measurement period with at least one eligible encounter during the measurement period

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).

Within the eMeasure, the denominator is defined as the initial patient population, which the specification defines as: "Patient Characteristic Birthdate: birth date" >= 12year(s) starts before start of "Measurement Period" AND: "Occurrence A of Encounter, Performed: Depression Screening Denominator Encounter Codes" (See attached code table for S2.b for specific value set codes included)

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients with an active diagnosis for Depression or a diagnosis of Bipolar Disorder are excluded.
Patients with any of the following are excepted: patient reason(s), patient refuses to participate, or medical reason(s); patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient´s health status; or situations where the patient´s functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools (for example: certain court appointed cases or cases of delirium).

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.)*
Within the eMeasure specification, value sets contain relevant codes to capture the exclusions. (See attached code table for S2.b for specific coding). The specification defines denominator exclusions as:
OR "Diagnosis: Depression diagnosis" satisfies all:
- starts before start of ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")
- overlaps ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")

OR "Diagnosis: Bipolor diagnosis" satisfies all:
- starts before start of ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")
- overlaps ("Encounter, Performed: Depression Screening Encounter Codes" during "Measurement Period")

The specification defines denominator exceptions as:
OR:
- AND: Union of:
  - "Risk Category Assessment not done: Medical or Other reason not done" for "Adolescent Depression Screening"
  - "Risk Category Assessment not done: Patient Reason refused" for "Adolescent Depression Screening"
  - during "Encounter, Performed: Depression Screening Encounter Codes"
- AND NOT: "Risk Category Assessment: Adolescent Depression Screening" during "Measurement Period"

OR:
- AND: Union of:
  - "Risk Category Assessment not done: Medical or Other reason not done" for "Adult Depression Screening"
  - "Risk Category Assessment not done: Patient Reason refused" for "Adult Depression Screening"
  - during "Encounter, Performed: Depression Screening Encounter Codes"
- AND NOT: "Risk Category Assessment: Adult Depression Screening" during "Measurement Period"

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.)*
No stratification.

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.)*
eMeasure PERFORMANCE CALCULATION –

To calculate provider performance, complete a fraction with the following measure components: Numerator (A), Performance Denominator (PD), Denominator Exclusions (B) and Denominator Exceptions (C).

Numerator (A): Number of patients meeting numerator criteria

Performance Denominator (PD): Number of patients meeting criteria for denominator inclusion

Denominator Exclusions (B): Number of patients with valid exclusions

Denominator Exceptions (C): Number of patients with valid exceptions.

1) Identify the patients who meet the eligibility criteria for the denominator (PD) which includes patients who are 12 years and older with appropriate encounters as defined by encounter codes or encounter value set during the reporting period.

2) Determine whether a Denominator Exclusion (B) applies and subtract those patients from the denominator.

3) Identify which of those patients meet the numerator criteria (A)

4) For those patients who do not meet the numerator criteria, determine whether an appropriate Denominator Exception (C) applies and subtract those patients from denominator (PD).

[Numerator (A) /[Performance Denominator (PD) - Denominator Exclusions (B) – Denominator Exceptions (C)]

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

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.)*
IF a PRO-PM, 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.
Electronic Health Record (Only)

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
No specific data source/data collection instrument.

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)*
Clinician Office/Clinic
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 a composite.

2. Validity – See attached Measure Testing Submission Form
Testing_form_NQF_0418-3132_CMS_2.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. (Do not remove prior testing information – include date of new information in red.)
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. (Do not remove prior testing information – include date of new information in red.)
No

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0418/3132
Measure Title: Preventive Care and Screening: Screening for Depression and Follow-Up Plan
Date of Submission: 12/9/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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*).

No existing dataset was used.

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

We tested the measure using electronic health record data from two practices for encounters from 1/1/2015 to 12/31/2015.

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.26*)**	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <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 recruited a primary care practice and a pediatrics practice in Pennsylvania that had the ability to report the measure and used two different EHR systems. Combined, the data from these two sites reflect 57 eligible professionals (EPs), each with an average of 953 patients.

Practice	Specialty	EHR Vendor	# of Providers	# Patients
A	Primary care	GE Centricity 12.0	53	52,961
B	Pediatrics	Medent 22.0	4	1,388

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*)

As shown in the previous section, two practices provided an extract of their EHR data containing encounter-level data for 54,349 eligible patients. The table below displays the distribution of patients across practices by sex, age strata, race, and ethnicity for the full EHR extract for each of the two sites.

	Practice A	Practice B	Total
Age			
12-17	870	1,087	1,957 (3.6%)
18-64	35,505	301	35,806 (65.9%)
65+	16,586	0	16,586 (30.5%)
Sex			
Male	24,342	711	25,053 (46.1%)
Female	28,605	677	29,282 (53.9%)
Unknown	14	0	14 (0%)
Race			
American Indian/Alaska Native	-	-	13 (0%)
Asian	-	-	327 (0.6%)
Black	342	101	443 (0.8%)
Native Hawaiian or Pacific Islander	-	-	38 (0.1%)
White	46,927	1,248	48,175 (88.6%)
Multiracial	-	-	18 (0.0%)
Unknown	5,322	13	5,335 (9.8%)
Ethnicity			
Hispanic or Latino	339	15	354 (0.7%)
Not Hispanic or Latino	49,073	1,353	50,426 (92.8%)
Unknown	3,549	20	3,569 (6.6%)

Results where value is less than or equal to 11 are not shown.

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.

EP performance scores and performance score reliability relied on data from all patients contained within the data extracted from the two sites' EHRs. We also conducted a systematic assessment of face validity, and supplement our findings with Bonnie testing results.

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

We aggregated performance scores calculated using EHR data by race, ethnicity, sex, and age to look for disparities. EHR data do not include information about income or other sociodemographic information.

2a2. RELIABILITY TESTING

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

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

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

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

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

We calculated reliability using a widely accepted method that is outlined in J.L. Adams' (2009) technical report titled "The Reliability of Provider Profiling: A Tutorial." In this context, reliability represents a measure's ability to confidently distinguish the performance of one physician from another. As discussed in the report, "Conceptually, [this method assesses] the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error." In this method, reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within providers), whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities. Although, there is not a clear cut-off for minimum reliability level, values above 0.7 are considered sufficient to see differences between physicians (or practices) and the mean, and values above 0.9 are considered sufficient to see differences between individual physicians or practices.

Adams, J.L. (2009). *The reliability of provider profiling: A tutorial* (TR-653-NCQA). Santa Monica, CA: RAND Corporation. Retrieved November 14, 2016, from http://www.rand.org/pubs/technical_reports/TR653.html

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*)

Performance measure score reliability (data combined across two sites):

Data source	Number of providers	Between-provider variance	Reliability mean	Reliability median	Reliability Std dev	Reliability min/max
EHR	52	.028	0.984	0.995	0.045	0.724 – 1.000

Note: Four providers were dropped from the reliability analysis who had 20 or fewer eligible patients.

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?*)

With average reliability score of 0.98, this measure demonstrates a high level of reliability to detect real difference in performance scores.

2b2. VALIDITY TESTING

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

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

☒ **Performance measure score**

☐ **Empirical validity testing**

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

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

Critical data elements (Bonnie testing):

We created a test deck consisting of 22 synthetic patient records to assess validity of the measure logic.

Sytematic assessment of face validity:

We surveyed 12 clinicians eligible to report this measure—none of whom advised on measure development—to rate face validity. We provided measure specifications and asked them to rate their agreement with the following statement: "The performance scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."

The rating scale offered five options: 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; 5 = strongly agree.

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

Critical data elements (Bonnie testing):

The expectations set for all 22 test cases were consistent with the actual values calculated by Bonnie (all cases "Pass"). The combined test deck utilizes 100% of the measure's logic statements and data elements.

Sytematic assessment of face validity:
1 – Strongly disagree – 0 votes
2 – Disagree – 3 votes
3 – Neither agree nor disagree – 0 votes
4 – Agree – 6 votes
5 – Strongly agree – 3 votes

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

Critical data elements (Bonnie testing):

The Bonnie test deck indicates that the measure logic accurately encodes the intent of the measure (see Appendix A for test deck details).

Sytematic assessment of face validity:

Nine of 12 experts (75 percent) agreed or strongly agreed that the measure accurately reflects quality. Experts who disagreed raised concerns related to patient compliance, documentation burden, and a preference that the measure specify one screening tool for adolescents, rather than several tools from which providers can choose.

_____

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We tested the frequency of exclusions using EHR data extracts from the two tesing sites.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The rate of exclusions and exceptions was 12.1% for all patients reported by the two practices that participated in testing the measure. The vast majority of these were patients identified as having been previously diagnosed with depression.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The exclusion rate is consistent with research on lifetime prevalence for depression and justifies the use of the exclusions/exceptions to account for situations in which it is appropriate not to screen and follow-up with patients for depression. Without these exclusions, measure performance could be skewed for EPs wth significant numbers of patients that were previously diagnosed with depression.

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

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

2b4.1.1 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

2b4.2. 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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

N/A

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

N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

N/A

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

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

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

N/A

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

N/A

2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
N/A

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

1b)

We used EHR data from two practices to calculate measure performance scores and assess the distribution of performance using statistical measures of central tendency (mean and median), variation (standard deviation), and spread (interquartile range and rates by decile).

We calculated chi-squared statistics to test for significant differences between expected and observed performance scores for various populations based on patients' race, ethnicity, sex, and age. Practices were not able to provide data on patients' socioeconomic status and/or disability status. These results represent only those providers who participated in the testing of this measure may not be generalizable to the population of all eligible providers.

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

EHR data from all two practices (data from 1/1/2015 through 12/31/2015):
Distribution of provider scores:

- Number of Providers 57
- Number of patients 54,349
- Average Unweighted Score 70.7%
- Average Weighted Score 68.3%
- Standard deviation 20.2%
- Interquartile range 18.2%
- Minimum 0.0%
- 10th percentile 40.9%
- 20th percentile 61.5%
- 30th percentile 67.5%
- 40th percentile 70.0%
- Median 72.6%
- 60th percentile 77.0%
- 70th percentile 79.4%
- 80th percentile 86.6%
- 90th percentile 93.8%
- Maximum 100.0%

Average performance score, by practice

Practice	Dates of data	Number of providers	Average weighted score	Average unweighted score
A	1/1/2015-12/31/2015	53	68.2%	70.7%
B	1/1/2015-12/31/2015	4	71.1%	70.5%

Please note: The unweighted average measure is the aggregated score for entire population. The weighted average is the average provider-level score, which is weighted by the number of patients in the denominator of each provider's score. All other statistics are based on weighted provider-level scores.

Proportion of denominator-qualifying patients receiving depression screening and appropriate follow-up by demographic category, 2015 EHR data:

Age Groups
12–17: 53.7%
18–64: 58.3%
65+: 91.4%
(X^2 = 5,252.569; df = 2; N = 47,782; p < 0.0001)

Race
American Indian or Alaska Native: 46.2%
Asian: 52.9%
Black: 72.4%
Native Hawaiian or other Pacific Islander: 51.4%
White: 69.4%
Multiracial: 72.2%
Unknown: 58.7%
(X^2 = 270.069; df = 6; N = 47,782; p < 0.0001)

Ethnicity
Hispanic or Latino: 59.6%
Not Hispanic or Latino: 68.4%
Unknown: 66.5%
(X^2 = 15.823; df = 2; N = 47,782; p = 0.0004)

Sex
Female: 68.5%
Male: 68.0%
(X^2 = 1.362; df = 1; N = 47,768; p = 0.2431)
We excluded 14 patients whose sex was unknown

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

Reported performance rates indicate a wide degree of variation and nearly all EPs in the two participating practices have the potential to improve the rates of depression screening and follow-up. The differences in performance rates by age groups is statistically significant and large enough to suggest potential clinical significance in the population studied, at least when comparing older patients (65 years or older) to younger ones. Quality improvement efforts should attempt to address these disparities. The differences by race and ethnic groups are also statistically significant, but the magnitude of the observed disparities may not be clinically significant, especially since there are small numbers of patients in some racial and ethnic groups. Differences in rates between males and females are not statistically significant. We did not stratify the measure based on age, race, or ethnicity because: (1) many other process measures show similar racial disparities and (2) stratifying the measure would significantly complicate implementation, reporting, and interpretation. The results of this analysis reflect a convenience sample of two practices and one of them is a pediatric practice, so it may not be representative of all EPs or group practices.

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

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

Claims, registry, and eCQM specifications are aligned across reporting methods. As directed by NQF, claims and registry testing data are submitted as NQF0418-3148.

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Entities report the measure using a single data source. We did not compare performance rates between the claims/registry measure and the eCQM because the claims and registry version of the measure is submitted separately. However, we have designed the specifications for all the data sources to maximize alignment and consistency.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

_____

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

The data we received from two practices was electronically extracted from each site's EHR, and both reports displayed patient-level data including age, sex, race, ethnicity, screening results, follow-up interventions, and presence of exclusion or exception critieria. Practice A did not provide data for one type of follow-up intervention (suicide risk assessment) or for denominator exception variables (medical reason, patient refusal). Discussions with the practice suggest that suicide assessments were captured in the "additional evaluation" field. This, combined with the fact that exceptions are optional variables, suggests that the absence of those data should not bias the performance score. Practice B also did not provide data for exception variables, nor did it report any of the following follow-up interventions: referral, additional evaluation, and suicide risk assessment. Similar to Practice A, discussions with staff at Practice B suggest that follow-up interventions were documented in a the "follow-up" variable, so lack of data in other follow-up variables should not bias performance scores.

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

We received a total of 54,349 patient records from the two practices which provided all 2015 patient encounters. As noted above, we did not receive data for all of the measure's data elements, but do not believe this biased performance scores. Additionally 14 (<1%) of patients did not have a reported sex, 67 patients (<1%) did not have a reported provider. We dropped patients without a known provider from provider-level results, and did not include patients with a reported sex in the disparities analysis. Given the low frequency, dropping these data did not bias performance scores. We did include patients whose race and/or ethnicity were unknown in the analysis of disparities in order to observe whether performance among these patients was meaningfully different than for the rest of the population. Results among patients of unknown ethnicity were comparable to those of patient not Hispanic or Latino.

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

As noted above, data from missing follow-up variables were captured within other follow-up variables, so we do not believe these missing data bias performance scores. Missing denominator exception data may impact performance scores, but given that these variables are optional, we believe our testing results are an accurate reflection EP experience calculating and reporting the eCQM. Although some patients were missing data related to sex and provider, the low frequency of these missing data should not impact performance scores. Results among patients of unknown race, which made up nearly 10 percent of patients, were lower than average but higher than rates among Asian, American Indian or Alaska Native, and Native Hawaiian or other Pacific Islanders.

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, Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.
ALL data elements are in defined fields in 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).
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: Feasibility_Scorecard_NQF_0418-3132.pdf

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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
We conducted workflow assessments at three practices and worked with these practices to complete feasibility scorecards. Two practices experienced feasibility challenges with data elements identifying follow-up interventions, particularly “additional evaluation,” “follow-up for depression,” and “suicide risk assessment;” but the third practice was able to capture these data elements. All three practices faced challenges with at least one of the data elements identifying denominator exceptions: “medical reason contraindicated” and “patient refused.” Since these are likely to be relatively infrequent, however, we do not expect these challenges to significantly impact measure performance. Overall, the measure was feasible in each of the three provider practices that contributed to testing of this eCQM. Most data elements necessary to capture the measure were regularly captured in structured fields.

This measure is not a PRO-PM.

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)*.
None

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 Physician Quality-Reporting System http://www.cms.gov/PQRS

4a.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

The Centers for Medicare & Medicaid Services (CMS) sponsors the Medicare and Medicaid EHR Incentive Programs (commonly referred to, collectively, as the Meaningful Use program) which provide financial incentives to entities that leverage certified EHR technology to improve patient care as outlined in specific program objectives. Eligible professionals (EPs), eligible hospitals, and critical access hospitals are required to report electronic clinical quality measures (eCQMs) during each year of participation in order to receive an incentive payment. More information about the Meaningful Use program is available at http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Meaningful_Use.html. At this time, no publicly available data are available on the frequency with which this measure is reported as part of the Meaningful Use Program.

The Physician Quality Reporting System (PQRS), also sponsored by CMS, is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by EPs. To be eligible for an incentive payment, EPs must satisfactorily report data on quality measures for covered Physician Fee Schedule services furnished to Medicare Part B Fee-for-Service beneficiaries. More information about PQRS is available at http://www.cms.gov/PQRS. According to the 2014 PQRS Reporting Experience, in 2014, this measure was one of six program measures in which more than 500,000 professionals were eligible to report, yet only 7.5 percent of those eligible actually reported. EP performance scores that rely on registry reporting are posted on Physician Compare. Individual eligible EPs and group practices participating under the PQRS Group Practice Reporting Option (GPRO) may report electronically using an EHR, and these results are also posted on Physician Compare.

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

N/A

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

4b. **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.
Average PQRS reporting rates from 2011 to 2014 reflect reporting by all participating providers, including those who reported the measure using EHR, claims, and registry data. EPs submit performance data voluntarily, and results may not be representative of all EPs. We do not have access to data on historical trends in performance specific to eCQM reporting, nor on performance rates by geographic area.

The average performance rate based on all data sources has fluctuated substantially over the past four years, decreasing from 82.6 percent in 2011 to 52.4 percent in 2014. However, the number of EPs reporting the measure has increased significantly over this time frame, from just 0.6 percent of EPs in 2011 to 7.5 percent in 2014. This makes it difficult to assess trends over time, as the EPs who recently began reporting the measure may have lower performance rates than those who have been reporting it for a longer period. Although the reporting increased each year, a substantial number of EPs are still not reporting the measure, and the average performance rate illustrates that there is still a gap in care.

4c. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
We have not identified any unintended consequences in our recent testing, or in the measure's implementation.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
We have not identified any unexpected benefits in our recent testing, or in the measure's implementation.

4d1.1. **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

4d1.2. **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

4d2.1. **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

4d2.2. **Summarize the feedback obtained from those being measured.**
N/A

4d2.3. **Summarize the feedback obtained from other users**
N/A

4d.3. **Describe how the feedback described in 4d.2 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

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)**
0518 : Depression Assessment Conducted

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
There are no competing measures. Multiple related measures have lost their NQF endorsement, including:
•Percent of Residents Who Have Depressive Symptoms (Long-Stay) – Centers for Medicare & Medicaid Services (formerly NQF #0690)
•Depression Screening by 13 Years of Age – National Committee for Quality Assurance (formerly NQF #1394)
•Maternal Depression Screening – National Committee for Quality Assurance (formerly NQF #1401)
•Depression Screening by 18 Years of Age – National Committee for Quality Assurance (formerly NQF #1515)

We also identified the following measures in the National Quality Measures Clearinghouse that do not have NQF endorsement:
•Adult depression in primary care: percentage of perinatal patients with documentation of screening for major depression or persistent depressive disorder using either PHQ-2 or PHQ-9 (Institute for Clinical Systems Improvement [ICSI])
•Adult depression in primary care: percentage of patients with cardiovascular disease with documentation of screening for major depression or persistent depressive disorder using either PHQ-2 or PHQ-9 (ICSI)
•Adult depression in primary care: percentage of patients who had a stroke with documentation of screening for major depression or persistent depressive disorder using either PHQ-2 or PHQ-9 (ICSI)
•Pediatric preventive care: percentage of pediatric patients aged 12 to 17 years who have a documented mental health and/or depression screening using one of the specified validated tools at a well-child visit during the measurement period (Minnesota Community Measurement)

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 only related NQF endorsed measure identified is 0518: Depression Assessment Conducted. Measure 0518 is an episode-based measure and reported based on OASIS data specific to home health agencies. It is similar to 0418, as it assesses depression using a standardized tool, but it differs in two key ways: First, target population: the denominator incorporates only adults aged 18 years and older and includes the number of home health episodes of care ending during the reporting period. Second, measure focus: the measure focuses on home health care in which patients received screening for depression. It does not include any follow-up component. 0418 is a patient-based measure focused on patients 12 years and older and includes a follow-up plan for positive depression screening results. Both are process measures; however, data for 0518 are only reported electronically and 0418 data may be reported using claims, registry, and electronic sources. 0418 is more robust in that it includes a broader population and requires a follow-up plan of care.

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 that target the same measure focus and or 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.
Attachment **Attachment:** NQF_0418_Summary_Materials-636173293393040868.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Jennifer, Harris, Jennifer.Harris@cms.hhs.gov, 410-786-3855-
Co.3 **Measure Developer if different from Measure Steward:** Quality Insights of Pennsylvania
Co.4 **Point of Contact:** Anita, Somplasky, asomplasky@wvmi.org, 877-346-6180-7852

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.
Through a collaborative process, the expert workgroup annually reviews the measure specifications (description, numerator, denominator, definitions, and clinical recommendation); literature review findings; and feedback or questions about the measure during its implementation. When last convened in 2016, the expert workgroup included the following members:

Jean Carter, PhD

Psychology
Washington Psychological Center, P.C.

Paula Hartman-Stein, PhD
Clinical psychology
Center for Healthy Aging; clinical psychologist, founder

Bracken Babula, MD
Internal medicine
Department of Medicine; Thomas Jefferson University; associate quality officer

Alan Axelson, MD
Adolescent psychiatry
InterCare Psychiatric Services; medical director and chief

Justin Schreiber, DO, MPH
Psychiatry
Western Psychiatric Institute and Clinic; co-triple board chief

Gregory M. Martino, PhD
Clinical psychology
Independent practice, DuBois, Pennsylvania

Tracy Murphy, AuD
Audiology
North Shore Audio-Vestibular Lab

Virginia Clark, PhD
Psychology (adolescent)
Western Reserve Psychological Associates, Inc.; president

Donald Wilson, MD
Obstetrics/gynecology
Women's Care Florida; chief medical officer

Harold Manley, PharmD
Pharmacology
Dialysis Clinic, Incorporated; director of medication management and pharmacovigilence

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2008

Ad.3 **Month and Year of most recent revision:** 04, 2016

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?** 04, 2017

Due to technical limitations, registered trademarks are indicated by (R) or [R] and unregistered trademarks are indicated by (TM) or [TM].

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:** 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 #: 0418:3148
Corresponding Measures: 0418:3132
Measure Title: Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: Percentage of patients aged 12 years and older screened for clinical depression on the date of the encounter using an age appropriate standardized depression screening tool AND if positive, a follow-up plan is documented on the date of the positive screen.
Developer Rationale: This measure aligns with the U.S. Preventive Services Task Force's (USPSTF) guidelines recommending routine screening for depression as a part of primary care for both children and adults, seeking to increase detection and treatment of depression and reduce the associated economic burden. The measure is an important contribution to the quality domain of community and population health.
The World Health Organization describes major depression as the leading cause of disability worldwide (Pratt & Brody, 2008). According to the Center for Behavioral Health Statistics and Quality (2015), in 2014, 11.7 percent of adolescents aged 12 to 17 and 6.6 percent of adults 18 years and older in the United States received a diagnosis of major depressive disorder. A study by Borner et al. (2010) found that 20 percent of adolescents are likely to have experienced depression by the time they are 18 years old. In adults, depression is the leading cause of disability in high-income countries and is associated with increased mortality due to suicide and impaired ability to manage other health-related issues (Siu, 2016).
The effects of depression in adults can include difficulties in functioning at home, in the workplace, and in social situations (Pratt & Brody, 2008). For example, 35 percent of men and 22 percent of women with depression reported that their depressive symptoms make it difficult for them to work, accomplish tasks at home, or get along with other people (Pratt & Brody, 2008). Effects of depression in adolescents are similar to those in adults; however, Siu (2016) noted depression has a negative effect on developmental trajectories in children and adolescents younger than 18 years old. Also, major depressive disorder in the adolescent population is especially problematic because it is linked with higher possibility of suicide attempt, death by suicide, and recurrence of the disorder in young adulthood.
Evidence strongly recommends screening for depression in adolescent and adult patients. Specifically, the USPSTF found convincing evidence that screening in primary care settings improves accurate identification of adolescent and adult patients with depression (Siu, 2016). Yet Borner et al. (2010) cite evidence that physicians are identifying and treating depression among adolescents even less than among adults, and that more than "70 percent of children and adolescents suffering from serious mood disorders go unrecognized or inadequately treated" (Borner, 2010, p. 948). Additionally, according to the 2016 USPSTF guideline for screening for depression in children and adolescents, only 36 to 44 percent of children and adolescents with depression receive treatment, further evidence that the majority of depressed children and adolescents go untreated. Although primary care providers (PCPs) are the first line of defense in detecting depression, studies show that PCPs fail to identify up to 50 percent of depressed patients, due to both lack of time and a lack of brief, sensitive, and easy-to administer psychiatric screening tools (Borner, 2010).

Finally, according to the 2016 USPSTF guideline for screening depression among adults, the United States spent about $22.8 billion on depression treatment in 2009, and an additional estimated $23 billion on lost productivity (Siu, 2016). This substantial economic burden warrants regular screening for depression, as screening is the first step in identifying those at risk for developing major depressive disorder and closing the performance gap.

Numerator Statement: Patients screened for clinical depression on the date of the encounter using an age appropriate standardized tool AND, if positive, a follow-up plan is documented on the date of the positive screen
Denominator Statement: All patients aged 12 years and older
Denominator Exclusions: Not Eligible – A patient is not eligible if one or more of the following conditions are documented:
•Patient refuses to participate
•Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status
•Situations where the patient's functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools. For example: certain court appointed cases or cases of delirium
•Patient has an active diagnosis of Depression
•Patient has a diagnosed Bipolar Disorder

Measure Type: Process
Data Source: Claims (Only), Registry
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

Original Endorsement Date: Jul 31, 2008 **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 *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.
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

- This measure was previously evaluated as measure #0418.
- The developer cited several individual studies, literature reviews, and a consensus statement that supported use of various screening instruments, made recommendations about treatment for children and adolescents with mood disorders, provided evidence of gaps in care for post-partum women systematic reviews, and provided estimates of prevalence of depression in the U.S. They also cited a 2009 USPSTF recommendation statement for screening for depression in adults, a 2009 USPSTF systematic review on screening for depression children and adolescents in the primary care setting, a 2010 AHRQ/USPSTF clinical practice guideline recommendation for

screening of children and adolescents, and two Institute for Clinical Systems Improvement (ICSI) clinical practice guideline recommendations for screening of adults (2012) and children and adolescents (2011).

- The Committee agreed that expansion of the measure to include patients 12 to 17 years of age is supported by the USPSTF recommendation, although they noted that some primary care providers may not be able to ensure accurate diagnosis, psychotherapy, and follow-up and therefore may not screen adolescents.
- The Committee discussed, at length, the measure's requirement for annual screening for depression, noting that the USPSTF recommendation does not specify screening intervals and the other guidelines recommend different intervals. Ultimately the Committee reached consensus supporting annual screening, agreeing the potential benefits of screening outweighed the risks.
- The Committee suggested that in the future, the developer should extend the measure to consider not just whether a follow-up plan is in place, but also whether the follow-up plan is implemented.

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:

- In their logic model, the developers link screening for depression to receipt of follow-up care, which they then link to improved health and quality of life.
- 2016 USPSTF recommendation statement on screening for depression in children and adolescents (moderate quality evidence, "B" recommendation).
- 2016 USPSTF recommendation statement on screening for depression in adults (moderate quality evidence, "B" recommendation).
- 2016 ICSI guideline recommendations for treating depression in adults in the primary care setting (low quality evidence, strong recommendation for low quality evidence)

Exception to evidence: N/A

Questions for the Committee:

- *The evidence provided by the developer is updated but directionally the same as 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 and grading (Box 3) → QQC presented (Box 4) → Quantity: high; Quality: low to moderate; Consistency: moderate (Box 5) → Moderate (Box 5b) → Moderate

The highest possible rating is 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 provided annual average performance rates for 2011-2014, based on data from the PQRS program, as well as CY 2015 distributional statistics from both claims and registry data

PQRS data, 2011-2014

Year	Average performance	Percent of eligible professionals reporting
2011	82.6%	0.6%
2012	65.2%	0.4%
2013	71.0%	1.3%
2014	52.4%	7.5%

Calendar year 2015 data

Source	# providers	# cases	Mean	Standard deviation	10th percentile	30th percentile	Median	70th percentile
Claims	26,169	3,002,169	36.5%	45.9%	0%	5.9%	100%	100%
Registry (provider)	7,027	989,092	28.9%	44.3%	0%	2.2%	50.8%	100%
Registry (practice)	1,797	989,092	***Unclear***	41.8	0%	12.1%	63.4%	99.2%

Disparities

- The developer provided patient-level data from 2015 Medicare claims. These results indicate that prevalence of screening varies according to age group, race, and sex, with lower rates in younger patients, blacks and whites, and males.
- The developer also cited recent literature indicating lower rates of screening and treatment in minority adults and lower rates of screening among men.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that disparities exist [for screening for depression screening and documenting a follow-up plan] in other patient subpopulations?*

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 to Support Measure Focus

Comments:

**agree with importance of screening for depression and appropriate assessment/treatment/referral. should be looking at appropriate training for those in primary care/med students/residents.
**Other than the interval for screening, the evidence, causal pathway and importance of this measure remain substantively the same.
**intermediate outcomes is the screening and outcome of interest is depression treatment or resolution. There is good evidence that depression can be identified, but only moderate evidence that identification leads to effective treatment (for non integrated primary care systems). There is very good evidence that it is beneficial for those whose depression are effectively treated. No new information that would change the evidence base for this measure. Yes, likely no need to repeat discussion to vote on evidence. Agree with algorithm outcome of moderate for evidence rating.
**Evidence is Grade 1 from a systematic review. The process of screening leads to early detection and intervention, which ultimately improves outcomes.

1b. Performance Gap

Comments:

**No additional comments.
**Continues to be a gap as demonstrated.
**yes, the percentage of providers who meet this measure is quite low nationally, suggesting ample room for improvement. Yes, gaps exist by age, minority status, and sex. Agree with "high" rating for improvement opportunity.
**There is much variation in performance and less than optimal performance with many disparities across population groups. there is much opportunity for improvement.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Claims and registries
Specifications:

- This measure is specified for the individual clinician and clinical group/practice levels of analysis in the clinician office/clinic setting. A higher score indicates better quality.
- The numerator, defined through use of G-codes, includes the number of patients screened using an age-appropriate standardized tool and, if positive, those for whom a follow-up plan is documented at the time of the positive screen.
    - A listing of examples of screening tools that would be appropriate for meeting the measure is provided.
    - Documentation of a follow-up plan must include at least one of the following: additional evaluation for depression; Suicide Risk Assessment; referral to a practitioner who is qualified to diagnose and treat depression; pharmacological interventions; or other interventions or follow-up for the diagnosis or treatment of depression.
- The denominator includes patients age 12 or older with a clinician office/clinic encounter (defined by CPT codes and G-codes).
- Exclusions include encounters where: the patient refuses to participate; the patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status; situations where the patient's functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools; patient has an active diagnosis of depression; or patient has a diagnosed bipolar disorder.
- A straightforward calculation algorithm is provided, which should allow for consistent calculation of the measure.
- The developer has indicated that only minor coding changes have been made to the specifications since the last NQF evaluation of the measure.
- In section 2b7, the developer notes that when testing the measure, they excluded claims without any of the relevant G-codes for the numerator. It is unclear if this is part of the specification of the measure or not.

Questions for the Committee:

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

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Inter-rater reliability testing of the data elements was conducted on a random sample of 275 Medicare claims.

Describe any updates to testing:

- Score-level testing using a signal-to-noise analysis was conducted using 2015 claims and registry data.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data element testing
  - Data used in testing included a random sample of 275 Medicare claims (from 77 randomly chosen providers) for the period Jan-Mar 2012.
  - Inter-rater reliability (IRR) between two nurse reviewers was calculated using percent agreement and kappa statistics. The nurses compared abstracted data from the medical records associated with the claims in the testing sample. IRR is an appropriate method for demonstrating data element reliability.
- Score-level testing
  - Reliability of the measure score was assessed using CY15 Part B Medicare claims data (n=26,169 providers, 3,002,169 patients) and PQRS registry data (n=7,027 individual providers, 1,727 practices, 989,092 patients).
  - A signal-to-noise analysis using the beta-binomial method was conducted. This type of analysis, which is an appropriate method for demonstrating score-level reliability, quantifies the amount of variation in performance that is due to differences between providers compared to differences due to measurement error. Results will vary based on the amount of variation between the providers and the number of patients treated by each provider. This method results in a reliability statistic that ranges from 0 to 1. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Results of reliability testing

- Data element testing
  - Numerator: 89.7% agreement, prevalence-adjusted kappa (PAK)=.80 (95% CI .70 -.89), Kappa=.75 (95% CI .64 - .86)
  - Denominator: 100% agreement
  - Denominator Exclusions: 66.5% agreement, PAK=.39 (95% CI .30 -.48), Kappa .18 (95% CI .09 -.27)
  - According to the Landis and Koch classification, a kappa value of 0.75 indicates substantial agreement, a kappa value of 0.18 indicates slight agreement, and a kappa value of 0.39 indicates fair agreement.
  - The developers attributed the lack of agreement between nurse reviewers for exclusions to use of different data sources to determine eligibility for exclusion and lack of specificity of the exclusion criteria language.
- Score-level testing

Data source	Number of providers/ practices	Between-provider variance	Reliability mean	Reliability median	Reliability standard deviation	Reliability min/max
Claims	26,169	.21	.99	1.0	.03	.62 - 1.0

Registry – provider level	7,027	.19	.99	1.0	.04	.60 - 1.0
Registry – practice level	1,797	.18	.99	1.0	.05	.59 - 1.0

Questions for the Committee:

- o *Are the test samples adequate to generalize for widespread implementation?*
- o *Do you have any concerns about the ability to code the measure exclusions consistently?*
- o *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing with measure as specified (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → High certainty that measure results are reliable (Box 6a) → High

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence: The evidence does not specify an optimal frequency for screening.

Question for the Committee:

- o *Given a lack of evidence regarding an optimal frequency for screening for depression, do you agree that screening at each visit is reasonable?*

2b2. Validity testing

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

For maintenance measures, summarize the validity testing from the prior review:

- Using a random sample of 275 Medicare claims, data element validity testing was conducted by comparing information on the claim to that included in the medical record.
- The developer alluded to the conduct of a face validity assessment but did not provide details regarding the process or results.

Describe any updates to validity testing:

- A face validity assessment was described, although it is unclear if this reflects a new assessment or additional details from the previous assessment.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

☐ **Empirical validity testing of the measure score**

Validity testing method:

- Data element testing
  - Data used in testing included a random sample of 275 Jan-Mar, 2012 Medicare claims (from 77 randomly chosen providers).
  - Developers calculated percent agreement statistics and kappa statistics were provided to indicate agreement between claims data and medical record data. Sensitivity and specificity statistics were not provided.
- Face validity testing
  - For the previous evaluation, no details on the face validity methodology were provided.
  - For the current evaluation, after providing the measure specifications, developers asked 12 clinicians who were eligible to report on the measure to rate their agreement regarding whether the measure results will provide an accurate reflection of quality and can be used to distinguish good and poor quality.
  - The clinicians included in the face validity assessment were not involved in the development of the measure.

Validity testing results:

- Data element testing
  - Numerator: 79.2% agreement, prevalence-adjusted kappa (PAK)=.60 (95% CI .51 -.69), Kappa=.38 (95% CI .25 - .50)
  - Denominator Exclusions: 93.0% agreement, PAK=.86 (95% CI .80 -.92), Kappa .64 (95% CI .49 -.79)
  - According to the Landis and Koch classification, a kappa value of 0.38 indicates fair agreement, a kappa value of 0.60 indicates moderate agreement, a kappa value of 0.64 indicates substantial agreement, and a kappa value of 0.86 indicates almost perfect agreement.
- Face validity testing
  - Of the 12 clinicians surveyed, 9 (75%) agreed or strongly agreed that the measure accurately reflects care quality, while 3 (25%) disagreed. Those who disagreed noted issues related to patient compliance, burden of documentation, and desire for use of only one specified tool for screening for adolescents.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Exclusions include encounters where: the patient refuses to participate; the patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status; situations where the patient's functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools; patient has an active diagnosis of depression; or patient has a diagnosed bipolar disorder.
- In CY 2015 Medicare claims, 3.6% of eligible encounters were excluded.
- In CY 2015 registry data, 4.9% of eligible encounters were excluded.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*

o *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

Source (CY2015)	# providers	# cases	Mean	Standard deviation	10th percentile	30th percentile	Median	70th percentile
Claims	26,169	3,002,169	36.5%	45.9%	0%	5.9%	100%	100%
Registry (provider)	7,027	989,092	28.9%	44.3%	0%	2.2%	50.8%	100%
Registry (practice)	1,797	989,092	*Not provided*	41.8	0%	12.1%	63.4%	99.2%

Question for the Committee:

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

2b6. Comparability of data sources/methods:

- No information on comparability between claims and registry data was provided.

2b7. Missing Data

- The developer noted that there was no missing information in their 2015 claims or registry data. However, they did state that claims without any of the G-codes used in the measure were excluded from the analysis.

Guidance from the Validity Algorithm

Specifications are mostly consistent with the evidence (frequency not addressed) (Box 1) → Threats to validity mostly assessed (no comparison of claims vs. registry results) → empirical validity testing for the measure as specified was conducted (Box 3) → No score-level empirical testing conducted (only face validity) (Box 6) → Data element level testing conducted (Box 10) → Method was reasonably appropriate (Box 11) → Testing results suggest moderate certainty (Box 12a) → Moderate

The highest possible rating is MODERATE (because score-level testing was not conducted).

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

Comments:

**No additional comments.
**Looks reasonable.
**most data elements are clearly defined, not sure I understand the numerator inclusion of G8432/8511 under S.5 because they seem to be incomplete screenings or denominator exclusions under S.9 though I'm guessing these two g codes represent cases where clinicians are documenting the patient as not eligible for some aspect of the measure. Algorithm seems clear. Yes, it's likely this algorithm could be consistently implemented.
**Data elements clearly defined in terms of whether screening did or did not occur; follow-up documentation may create some challenges.

2a2. Reliability Testing

Comments:
**No additional comments
**Yes, fine.
**Yes, adequate scope was used to generalize results. yes, there's substantial agreement, though it seems clarity around what constitutes an exclusion from denominator could be and should be improved. Yes, it appears the measure can reliably detect differences among providers/populations. High rating on reliability probably makes sense.
**Data element and score level reliability testing conducted. Reliability is high.

2b1. Validity Specifications
Comments:
**No additional comments.
**Again, the only issue is frequency and there are no good data I am aware of that directly address this issue and demonstrate improvements in patient oriented outcomes.
**Given nature of depression, variety of etiologies, stigma and other aspects of screening such as suicide, depression screening at each appointment is reasonable. I would also suggest that measure is consistent w/ evidence as opposed to "somewhat". Should only exclude those w/ depression or bipolar if they're in remission or maybe treatment as surrogate.
**Specifications are consistent with the evidence.

2b2. Validity Testing
Comments:
**No additional comments
**Yes, well supported.
**Yes, testing occurred w/ adequate scope. yes, the measure demonstrates adequate validity. Data element suggested fair validity w/ Kappa of 0.38 and face validity demonstrated 75% agreement of validity. The submitter did comment on changes that were to be made including coding of screening tool used that would potentially increase validity.
**Validity testing involved measure score with face validity and data element testing. Validity moderate.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**question how those who refuse to fill out screening tool get accounted for?
**No real problems.
**codes not containing G code were excluded, better understanding of coding practices are probably needed to know if this is a significant problem. "moderate" rating seems reasonable.
**Exclusions appropriate; risk adjustment none

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 required data elements are routinely collected and are available in electronic sources.
- The developer did not report any implementation challenges.
- The developer did not provide information on feasibility of reporting via claims versus registry.

Questions for the Committee:
- *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

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
**may be difficult to sort out documentation of recommendations and if follow through is successfully implemented.
**Yes, proven feasible, if a bit cumbersome.
**screening codes and positive or negative is regularly captured and assuming the treatment plan portion is also included in codes, though not clear how often systems use coding to full extent. My concern would be to better understand how coding system is utilized under real world conditions.
**Feasibility moderate to high.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- This measure is used the CMS Physician Quality Reporting System (PQRS) *[which is being phased out by 12/31/18 and is replaced by MIPS]*.
  - In 2014, 7.5% of >500,000 eligible professionals reported on the measure.
- The measure also is included in the following CMS programs:
  - Electronic Health Record Incentive Program: EPs
  - Medicare Shared Savings Program (MSSP)
  - Merit-Based Incentive Payment System (MIPS) Program
  - Physician Value-Based Payment Modifier (VBM) *[which is being phased out by 12/31/18 and is replaced by MIPS]*
  - Physician Feedback/Quality and Resource Use Reports (QRUR) *[which is being phased out by 12/31/18 and is replaced by MIPS]*
  - Physician Compare
  - Medicaid Adult Core Set

Improvement results:

PQRS data, 2011-2014 (includes data reported through EHRs, claims, and registries)

Year	Average performance	Percent of eligible professionals reporting
2011	82.6%	0.6%
2012	65.2%	0.4%
2013	71.0%	1.3%
2014	52.4%	7.5%

Unexpected findings (positive or negative) during implementation: None reported.

Potential harms: None reported.

Vetting of the measure: None reported.

Feedback:

- In 2012, the Measure Applications Partnership (MAP) supported the measure for inclusion in the clinician Meaningful Use program and also specifically referenced the measure as relevant to the dual eligible beneficiary population.
- In 2012, HHS included in measure in the CMS initial core set of measures for Medicaid-eligible adults.
- In 2014, the MAP supported the measure for inclusion in the Physician Compare and Value-Based Payment Modifier programs. The MAP also supported its inclusion in the End-Stage Renal Disease Quality Incentive Program, even though the measure is specified for individual clinicians/practices and not for dialysis facilities. The MAP Medicaid Task Force also supported its continued use in the Medicaid Adult Core Set.
- In 2015, the MAP Medicaid Task Force again supported the measure's continued use in the Medicaid Adult Core Set.

***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?*
- o *Has the measure been vetted in real-world settings by those being measure or others?*
- o *Does inclusion of the measure in the QRUR program meet the requirements for vetting by those being measured?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

**would like to see evidence of screening actually leading to improved outcomes. I think the topic of appropriate training of med students, residents, and those in primary care deserves ongoing attention.
**One wonders if there are not improvements in performance, is the measure useful... well, I am just being contrarian...
**measure can be used to increase identification and treatment of depression in primary care including system changes to improve services in these areas such as integration of mental health services into primary care. Yes it's being publicly reported and used in multiple CMS programs. Use of the measure can bring depression awareness, identification and treatment. Unintended consequences would be increased screening efforts and focus of workforce on depression vs other issue in primary care. Prevalence, early age of onset, disability and increased co-morbid disease burden justify depression identification and treatment as priority focus area. Yes, the measure's been vetted in real world settings, submitter denies significant complaints or feedback on the measure, no changes made.
**The measure is currently publicly reported and used in the CMS Physician Quality Reporting System.

Criterion 5: Related and Competing Measures

Competing measures:

0518: Depression Assessment Conducted

3132: Preventive Care and Screening: Screening for Depression and Follow-Up Plan

- #3132 is the eMeasure "version" of this measure. NQF will not ask the Committee to select a best-in-class measure.

Harmonization:

- 0518

- Measure #0518 is a measure that is specified for facility-level assessment in the home health setting and thus the Committee will not be asked to select a best-in-class measure.
- During the 2014 evaluation of the measure, the developers agreed that were opportunities for harmonization, including adding the home health setting to this measure (#3148) and adding a follow-up requirement to #0518. The Committee recommended that the developers pursue harmonizing the measures in the areas of care settings and follow-up.
- No change to the care setting for this measure (formerly #0418) has been made.

- 3132
  - It appears that this eMeasure "version" has been harmonized with #3148 to the extent possible and thus NQF will not ask the Committee to discuss harmonization.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because measure score validity testing has not been conducted.

Pre-meeting public and member comments

- None received.

Measure Number (*if previously endorsed*)**:** 0418 (3132/3148)

Measure Title: Preventive Care and Screening: Screening for Depression and Follow-Up Plan

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: None

Date of Submission: 12/9/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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.

Evidence to Support the Measure Focus

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

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

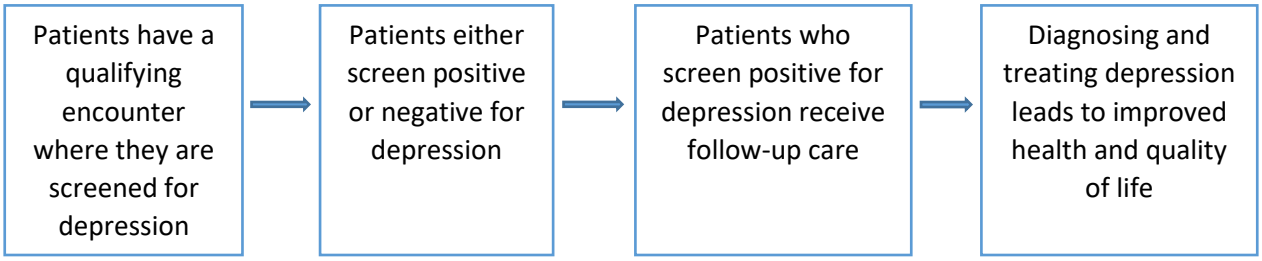
☒ Process: Depression screening and follow-up plan

☐ 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.12 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.



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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

N/A

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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**	• **Screening for Depression in Children and Adolescents: U.S. Preventive Services Task Force Recommendation Statement** • **Albert L. Siu, MD, MSPH, on behalf of the U.S. Preventive Services Task Force** • **Published February 9, 2016** • **U.S. Preventive Services Task Force. Screening for depression in children and adolescents: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2016 Mar 1;164(5):360-366** • **https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/depression-in-children-and-adolescents-screening1?ds=1&s=depression** **Previous Submission**: Williams, S.B., O´Connor, E.A., Eder, M., Whitlock, E.P. (2009). Screening for Child and Adolescent Depression in Primary Care Setting: A Systematic Evidence Review for the US Preventive Services Task Force. Pediatrics, 123, e716-e735. doi:10.1542/peds.2008-2415
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a	"The USPSTF recommends screening for major depressive disorder (MDD) in adolescents aged 12 to 18 years. Screening should be implemented with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up. (B recommendation)" (p. 360) **Previous Submission**: "The US Preventive Health Services Task Force recommends screening of adolescents (aged 12 to18 years) for (MDD) when systems are in place to ensure accurate diagnosis,

guideline, summarize the conclusions from the SR.	psychotherapy (cognitive behavioral or interpersonal), and follow-up (Grade B recommendation)" (AHRQ, 2010, p. 141)
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Moderate: "The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies; inconsistency of findings across individual studies; limited generalizability of findings to routine primary care practice; and lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion." (p. 367) **Previous Submission:** *Please see 1c.13 at end of document for grading of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Provide all other grades and definitions from the evidence grading system	As indicated in Appendix Table 2 (p. 360). High: "The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies." Low: "The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: the limited number or size of studies; important flaws in study design or methods; inconsistency of findings across individual studies; gaps in the chain of evidence; findings that are not generalizable to routine primary care practice; and a lack of information on important health outcomes. More information may allow an estimation of effects on health outcomes." **Previous Submission:** *Please see 1c.12 at end of document for grading of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Grade assigned to the **recommendation** with definition of the grade	B Recommendation: "The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial." (p. 367) **Previous Submission:** USPSTF Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.
Provide all other grades and definitions from the recommendation grading system	As indicated in Appendix Table 1 (p. 360). A: "The USPSTF recommends the service. There is high certainty that the net benefit is substantial." C: "The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small." D: "The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits." I statement: "The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor

	quality, or conflicting, and the balance of benefits and harms cannot be determined." **Previous Submission does not provide this information.**
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The USPSTF conducted a systematic evidence review to update its 2009 recommendation on screening for child and adolescent major depressive disorder (MDD) in primary care settings. Compared to its 2009 review, USPSTF narrowed the scope of this evidence review to focus exclusively on screening for and treating MDD. The USPSTF excluded studies of paroxetine because in 2003 the U.S. Food and Drug Administration (FDA) recommended not to use paroxetine to treat MDD in children and adolescents. The USPSTF examined evidence on the benefits and harms of screening; the accuracy of screening tests feasible for use in primary care; and the benefits and harms of treatment with psychotherapy, medications, and collaborative care models in patients ages 7 to 18 years. USPSTF limited treatment studies to those that were implemented in or required referrals from primary care settings to ensure that the patient population was similar to patients who would be identified through screening. The USPSTF found five good- or fair-quality studies of the accuracy of MDD screening instruments in children and adolescents ages 11 years or older. It also found eight good- or fair-quality randomized controlled trials (RCTs) that reported health outcomes in children or adolescents with screen-detected MDD: four for patients who were treated with selective serotonin reuptake inhibitors (SSRIs), two involving psychotherapy, one on SSRIs combined with psychotherapy, and one on collaborative care. Most trials were restricted to adolescent's ages 12 to 14 years or older; only two of the four SSRI trials included children ages 7 or 8 years (p. 364). **Previous Submission:** *Please see 1.c.5 and 1c.6 at end of document for body of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Estimates of benefit and consistency across studies	The USPSTF found adequate evidence that screening tests can accurately identify MDD in adolescents and that treatment of adolescents with screen-detected MDD is associated with beneficial reductions in symptoms. The USPSTF therefore concluded with moderate certainty that screening for MDD in adolescents ages 12 to 18 years is associated a moderate net benefit (p. 365). Although the USPSTF found no studies that directly evaluated whether screening for MDD in adolescents in primary care settings leads to improved health and other outcomes, there is adequate evidence that treatment of MDD detected through screening in adolescents is associated with moderate benefit, for example, by reducing the severity of depression or improving depression symptoms (p. 361). **Previous Submission:** *Please see 1c.7 at end of document for Estimates of benefit and consistency across studies for the USPSTF and ICSI guidelines which was provided in the previous submission
What harms were identified?	The USPSTF found no direct evidence to suggest that screening or treatment for MDD in adolescents or children leads to potential harms. Seven trials in the USPSTF review pertaining to the use of SSRIs (five trials), psychotherapy with or without

	SSRIs (one trial), or collaborative care (one trial) reported on harms, but none found significant differences between intervention groups. The USPSTF noted, however, that some of the studies had insufficient power to detect differences on some of the measured outcomes. **Previous Submission:** *Please see 1c.8 at end of document for harms identified for the USPSTF and ICSI guidelines which was provided in the previous submission.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	This guideline was published in 2016 and is the most recent systematic review completed. **The 2016 guideline is an update of the 2009 release, submitted in the previous submission.**

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Screening for Depression in Adults US Preventive Services Task Force Recommendation Statement** • **Albert L. Siu, MD, MSPH; and the US Preventive Services Task Force (USPSTF)** • **Published January 26, 2016** • **US Preventive Services Task Force (USPSTF). Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. JAMA. 2016; 315(4):380-387. doi:10.1001/jama.2015.18392.** • **https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/depression-in-adults-screening1?ds=1&s=depression** **Previous Submission:** U.S. Preventive Services Task Force (2009). Screening for Depression in Adults: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine, 151 (11), 784-792. Retrieved from: http://annals.org/article.aspx?articleid=745304 USPSTF Grade B and Grade C recommendation
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline,	"The USPSTF recommends screening for depression in the general adult population, including pregnant and postpartum women. Screening should be implemented with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up (B recommendation)" (p. 360). **Previous Submission:** The US Preventive Health Services Task Force (USPSTF) recommends screening adults for depression when staff-assisted depression care supports are in place to assure accurate diagnosis, effective treatment, and follow-up. Grade B Recommendation" (AHRQ, 2010, p. 136)

summarize the conclusions from the SR.	
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	As indicated in Appendix Table 2 (p. 367). Moderate: "The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: the number, size, or quality of individual studies; inconsistency of findings across individual studies; limited generalizability of findings to routine primary care practice; and lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion." **Previous Submission:** *Please see 1c.13 at end of document for grading of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Provide all other grades and definitions from the evidence grading system	As indicated in Appendix Table 2 (p. 367). High: "The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies." Low: "The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: the limited number or size of studies; important flaws in study design or methods; inconsistency of findings across individual studies; gaps in the chain of evidence; findings that are not generalizable to routine primary care practice; and a lack of information on important health outcomes. More information may allow an estimation of effects on health outcomes." **Previous Submission:** *Please see 1c.12 at end of document for grading of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Grade assigned to the **recommendation** with definition of the grade	As indicated in Appendix Table 1 (p. 367). Graded B Recommendation: "The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial." **Previous Submission:** USPSTF Grade B: The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.)
Provide all other grades and definitions from the recommendation grading system	As indicated in Appendix Table 1 (p. 367). A: "The USPSTF recommends the service. There is high certainty that the net benefit is substantial. Offer or provide this service." C: "The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgment and patient preferences. There is at least moderate certainty that the net benefit is small." D: "The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits."

	I statement: "The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined." **Previous Submission does not provide this information.**
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The USPSTF found convincing evidence that screening tests accurately identify depression in the general adult population, and that there are benefits to treating depression once diagnosed. Nine good- or fair-quality trials addressed screening in general adults and older adults (p.384). The evidence from five RCTs, in addition to indirect evidence reviewed for the 2009 recommendation, indicates that there is moderate certainty that screening for depression in adults is of moderate benefit (p. 385). In terms of treatment, two systematic reviews concluded that antidepressants were effective in treating depression in older adults. Two good-quality systematic reviews found that older adults who received psychotherapy were more than twice as likely to have remission as those who received no treatment (p. 384). For pregnant and postpartum women, the USPSTF reviewed 23 studies comparing the accuracy of the Edinburgh Postnatal Depression Scale with diagnostic interview, and found that the instrument had an acceptable positive predictive value for detecting MDD. The USPSTF identified six good- or fair-quality RCTs that assessed the effect of screening for depression in pregnant and postpartum women that support recommending depression screening for this group (p. 384). Eighteen trials examined the benefits of treatment interventions in women who screened positive for depression in primary care or community settings. Of these, 15 focused on postpartum women and 3 involved pregnant women. Ten RCTs found that cognitive behavioral therapy (CBT) benefits both postpartum and pregnant women; the remaining eight trials did not find sufficient evidence to draw conclusions about the effectiveness of treatment in these populations (p. 385). The USPSTF review found seven studies that compared suicide-related events in adults who received SSRIs and other antidepressants versus placebo. None of the studies showed a significant increase in suicide completion among adults taking antidepressants, but given the rarity of this event, they may not have had sufficient power to detect differences between treatment groups (p. 385). The majority of the evidence on the harms of antidepressants in pregnant and postpartum women comes from a good-quality comprehensive systematic review on the comparative effectiveness and safety of antidepressant treatment for depression in this population. The review, which included 124 observational studies, showed that second-generation antidepressant use during pregnancy may be associated with a small increase in the risk of several outcomes, including preeclampsia, miscarriage, and respiratory distress (p. 385). **Previous Submission:** *Please see 1.c.5 and 1c.6 at end of document for body of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Estimates of benefit and consistency across studies	The USPSTF concluded with at least moderate certainty that there is a moderate net benefit to screening for depression in adults, including older adults, and in pregnant and postpartum women who receive care in clinical practices that have CBT or other evidence-based counseling available after screening (p. 381).

	The USPSTF also found adequate evidence that programs that screen for depression and have adequate support systems in place improve clinical outcomes in both the general adult population and among pregnant and postpartum women specifically. Improvement in outcomes included reduction and remission of depression symptoms (p. 380). **Previous Submission:** *Please see 1c.7 at end of document for Estimates of benefit and consistency across studies for the USPSTF and ICSI guidelines which was provided in the previous submission
What harms were identified?	"The USPSTF found adequate evidence that the magnitude of harms of screening for depression in adults is small to none. The USPSTF found adequate evidence that the magnitude of harms of treatment with CBT in postpartum and pregnant women is small to none" (p. 380). "The USPSTF found that second-generation antidepressants (mostly selective serotonin reuptake inhibitors [SSRIs]) are associated with some harms, such as an increase in suicidal behaviors in adults aged 18 to 29 years and an increased risk of upper gastrointestinal bleeding in adults older than 70 years, with risk increasing with age; however, the magnitude of these risks is, on average, small. The USPSTF found evidence of potential serious fetal harms from pharmacologic treatment of depression in pregnant women, but the likelihood of these serious harms is low. Therefore, the USPSTF concludes that the overall magnitude of harms is small to moderate" (p. 381). **Previous Submission:** *Please see 1c.8 at end of document for harms identified for the USPSTF and ICSI guidelines which was provided in the previous submission.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	This guideline was published in 2016 and is the most recent systematic review completed. **The 2016 guideline is an update of the 2009 release, submitted in the previous submission.**

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number**	• **Health Care Guideline Adult Depression in Primary Care** • **Institute for Clinical Systems Improvement (ICSI)** • **March 2016** • **Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, Mack N, Myszkowski M. Institute for Clinical Systems Improvement. Adult Depression in Primary Care. Updated March 2016. Pp.1-131**

• **URL**	• **https://www.icsi.org/guidelines__more/catalog_guidelines_and_more/catalog_guidelines/catalog_behavioral_health_guidelines/depression/** **Previous Submission:** ICSI References below: Wilkinson, J., Bass, C., Diem, S., Gravley, A., Harvey, L., Hayes, R., Johnson, K., Maciosek, M., McKeon, K., Milteer, L., Morgan, J., Rothe, P., Snellman, L., Solberg, L., Storlie, C., Vincent, P. (2012). Institute for Clinical Systems Improvement. (18th ed.). Health care guideline: Preventive services for adults. Retrieved from: http://www.icsi.org/preventive_services_for_adults/preventive_services_for_adults_4.htm Wilkinson, J., Bass, C., Diem, S., Gravley, A., Harvey, L., Hayes, R., Johnson, K., Maciosek, M., McKeon, K., Milteer, L., Morgan, J., Rothe, P., Snellman, L., Solberg, L., Storlie, C., Vincent, P. (2011). Institute for Clinical Systems Improvement (17th ed.). Preventive Services for Children and Adolescents.
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.	All recommendations are contained within a table on pages 8-10. We have included below those that are most relevant to this measure: "Clinicians should routinely screen all adults for depression using a standardized instrument" (p. 8). "Clinicians should establish and maintain follow-up with patients" (p. 10). "Clinicians should screen and monitor depression in pregnant and post-partum women" (p. 10). **Previous Submission:** The Institute for Clinical Systems Improvement Preventive Services for Adults Health Care Guideline recommendation: "Routine depression screening should be recommended for adult patients (including older adults but only if the practice has staff-assisted 'systems in place to ensure that positive result are followed by accurate diagnosis, effective treatment and careful follow-up.' The optimum interval for rescreening is unknown (O'Connor, 2009 [Systematic Review])" (ICSI, 2012, p.25) The Institute for Clinical Systems Improvement Level II Services for Children and Adolescents recommendation: "Screen adolescents (ages 12-18) for major depressive disorder, but only when systems are in place for in their organization to ensure accurate diagnosis, careful selection of treatment and close follow-up" (ICSI, 2011, p.18).
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	All three of the recommendations listed above were graded as Low Quality Evidence: "Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain." (p. 4) **Previous Submission:** *Please see 1c.13 at end of document for grading of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Provide all other grades and definitions from	ICSI utilizes the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology system. Visit http://www.gradeworkinggroup.org/ for more information about GRADE.

the evidence grading system	High Quality Evidence: **"**Further research is very unlikely to change our confidence in the estimate of effect." (p. 4) Moderate Quality Evidence: "Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate." (p. 4) **Previous Submission:** *Please see 1c.12 at end of document for grading of evidence for the USPSTF and ICSI guidelines which was provided in the previous submission
Grade assigned to the **recommendation** with definition of the grade	Strong Recommendation for Low Quality Evidence: "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." (p. 4) **Previous Submission:** *Please see 1c.13 at end of document for grading of recommendation for the ICSI guidelines which was provided in the previous submission
Provide all other grades and definitions from the recommendation grading system	Strong Recommendation for High Quality Evidence: "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." (p. 4) Weak Recommendation for High Quality Evidence: "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." (p. 4) Strong Recommendation for Moderate Quality Evidence: "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." (p. 4) Weak Recommendation for Moderate Quality Evidence: "The work group recognizes that there is a balance between harms and benefits, 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." (p. 4) Weak Recommendation for Low Quality Evidence: "The work group recognizes that there is significant uncertainty about the best estimates of benefits and harms." (p. 4) **Previous Submission:** *Please see 1c.12 at end of document for grading of recommendation for the ICSI guidelines which was provided in the previous submission
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The authors used a consistent and defined literature search process to develop and revise the ICSI guidelines. First, ICSI staff, in consultation with the work group and a medical librarian, conducted a literature search to identify systematic reviews, randomized clinical trials, meta-analyses, other guidelines, regulatory statements, and any other pertinent literature. Work group members then evaluated the identified literature using the GRADE methodology (p. 131). For this guideline, ICSI reviewed 12 systematic reviews, 17 meta-analyses, 18 RCTs, 1 meta-regression, and 2 guidelines.

	The body of evidence related to screening all adults and pregnant and post-partum women was of low to moderate quality. The body of evidence for establishing a follow-up plan was of high quality. **Previous Submission:** *Please see 1.c.5 and 1c.6 at end of document for grading of recommendation for the ICSI guidelines which was provided in the previous submission
Estimates of benefit and consistency across studies	For the recommendation: "Clinicians should routinely screen all adults for depression using a standardized instrument," ICSI determined that screening results in finding and treating more depressed patients, leading to better outcomes and improved functioning not only for depression, but also for other co-morbid conditions. There is also some evidence that screening may reduce overall, long-term medical costs for depressed patients (p. 14). For the recommendation: "Clinicians should establish and maintain follow-up with patients," ICSI determined that appropriate, reliable follow-up is highly correlated with improved response and remission scores. Follow-up is also correlated with the improved safety and efficacy of medications and helps prevent relapse (p. 50). For the recommendation: "Clinicians should screen and monitor depression in pregnant and post-partum women," ICSI determined that screening patients leads clinicians to find and treat more patients with depression. Furthermore, untreated prenatal depression is associated with negative pregnancy outcomes such as poor maternal self-care, poor nutrition, preterm labor, and low birth weight. Untreated prenatal depression is also associated with negative effects on children such as developmental delay and cognitive impairment (p. 122). **Previous Submission:** *Please see 1c.7 at end of document for grading of recommendation for the ICSI guidelines which was provided in the previous submission
What harms were identified?	The only harms identified for all three recommendations were the cost of screening patients who are not depressed, and the potential additional cost of unnecessary follow-up visits (p. 14, p. 50, p. 122). **Previous Submission:** *Please see 1c.8 at end of document for grading of recommendation for the ICSI guidelines which was provided in the previous submission
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We have not identified any additional new studies published since the release of this guideline that would change its conclusions. The ICSI guideline is often based upon the USPSTF recommendations. The 2016 guideline is an update of the 2012 release, submitted in the previous submission The 2011 adolescent recommendations submitted in the previous submission are currently under review and being updated, but will likely be consistent with the USPSTF adolescent recommendations.

Previous Submission:

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade

1c.13 Grade Assigned to the Body of Evidence: Overall Evidence Grading: SORT Strength of Recommendation B: considerable patient-oriented evidence, i.e., re: improved recognition and diagnosis of depression, and improved depression outcomes, but not consistently high quality evidence. There is considerable and consistent patient-oriented research evidence, in addition to clinical recommendation statements, documenting the prevalence and burden of depression among adolescents, the importance of screening for depression among adolescents, and the availability of depression screening tools.

Provide all other grades and definitions from the evidence grading system

1c.12 If other, identify and describe the grading scale with definitions: The Strength of Recommendation Taxonomy (SORT)

An A-level recommendation is based on consistent and good-quality patient-oriented evidence; a B-level recommendation is based on inconsistent or limited-quality patient-oriented evidence; and a C-level recommendation is based on consensus, usual practice, opinion, disease oriented evidence, or case series for studies of diagnosis, treatment, prevention, or screening. The quality of individual studies is rated 1, 2, or 3; numbers are used to distinguish ratings of individual studies from the letters A, B, and C used to evaluate the strength of a recommendation based on a body of evidence.

Body of evidence:

- Quantity – how many studies?

1.c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles*):

Numerous studies were reviewed in the body of the evidence. Refer to 1c.6 for details of studies within evidence. Evidence is annually reviewed through an environmental scan of the measure focus and target population. The measure specification's rationale and clinical recommendation statements were reviewed and revised based on the current evidence found in the environmental scan.

For complete list of evidence with grading, refer to section 1c 15.

Quality – what type of studies?

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events*): The articles in the body of evidence review numerous studies addressing the measure focus and/or target population. They are listed in detail below:

• Borner et al. 2010 is one study examining the potential use of the Patient Health Questionnaire (PHQ-2) screening instrument for depression in adolescents when seen by their primary care physician.

• Coyle et al. 2003 is a consensus statement in which the expert panel reviewed 104 studies including many randomized controlled trials and other studies to make recommendations about risk factors, diagnosis, treatment and services in children and adolescents with mood disorders.

• Dunn et al. 2012 is one study which evaluated the psychometric properties of an adapted version of the Modified Depression Scale among 9th-12th graders in Boston through a school-based survey. This article also reviewed and compared at least 15 randomized controlled trial results, measurement tool literature and numerous review articles.

• Feinberg et al. 2009 is one study in which 42 women were screened for maternal depression and reported gaps in care for postpartum women.

• Liberto 2012 is an integrated review of 35 articles (which included both quantitative and qualitative research studies) addressing literature on screening for depression and help-seeking behaviors by postpartum women during pediatric well-baby visits; to identify gaps in the literature relating to depression and help-seeking behaviors; and to discuss implications for practice and future research.

• Pratt et al. 2008 is a data brief with data derived from the National Health and Nutrition Examination Survey, 2005-2006 The report did reference at least four randomized controlled trials.

• Williams et al. 2009 is a systematic evidence review by the U.S. Preventive Services Task Force which included 4 systematic reviews/meta-analyses and 31 studies (including fair to good quality randomized, controlled trials)

• Healthy People 2020 are a comprehensive framework of evidence-based objectives and goals targeted to improve the health behaviors of a nation which references numerous high quality data sets and randomized controlled trials.

• U.S. Preventive Services Task Force 2009 is a recommendation statement in which the researchers reported results of multiple randomized trials studies and meta-analyses for each aspect of diagnosis, treatment and follow-up as well as additional research-based studies.

The body of evidence consists of a total of thirteen evidences (excluding the three clinical guidelines listed in evidence guideline recommendation section 1c.16-1c.24). Two evidences have SORT Study quality level 1: good-quality patient-oriented evidence (Liberto, 2012 & Williams et al., 2009), One evidence has a USPSTF Grade B and Grade C Recommendation (USPSTF, 2009). Eight evidences have SORT Study quality level 2: limited-quality patient-oriented evidence (Borner et al., 2010; Centers for Disease Control and Prevention, 2007; Dunn et al., 2012; Feinberg et al., 2009; Geriatric Mental Health Foundation, 2008;Pratt & Brody, 2008; Steinman et al., 2007; Unützer et al., 2009). Two evidences have SORT Study quality level 3: other evidence: guideline (U.S. Department of Health and Human Services, 2011; Coyle et al., 2003). The evidence bears directly on the importance, benchmarking, performance gaps and disparities of depression screening and follow-up in the outpatient setting and the potential reduction of negative outcomes with improved recognition and diagnosis of depression, and improved depression outcomes. Since the studies show consistently statistically significant effects, there are no issues of "imprecision/wide confidence intervals due to few patients or events".

Estimates of benefit and consistency across studies

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* Consistency of results across studies: While the magnitude of the effects varies from study to study, the effects are consistently positive.

What harms were identified?

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

Studies show consistent benefits while detecting no harm and yielding consistent net benefits.

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.

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_form_NQF_0418-636178224651499436.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

This measure aligns with the U.S. Preventive Services Task Force's (USPSTF) guidelines recommending routine screening for depression as a part of primary care for both children and adults, seeking to increase detection and treatment of depression and reduce the associated economic burden. The measure is an important contribution to the quality domain of community and population health.

The World Health Organization describes major depression as the leading cause of disability worldwide (Pratt & Brody, 2008). According to the Center for Behavioral Health Statistics and Quality (2015), in 2014, 11.7 percent of adolescents aged 12 to 17 and 6.6 percent of adults 18 years and older in the United States received a diagnosis of major depressive disorder. A study by Borner et al. (2010) found that 20 percent of adolescents are likely to have experienced depression by the time they are 18 years

old. In adults, depression is the leading cause of disability in high-income countries and is associated with increased mortality due to suicide and impaired ability to manage other health-related issues (Siu, 2016).
The effects of depression in adults can include difficulties in functioning at home, in the workplace, and in social situations (Pratt & Brody, 2008). For example, 35 percent of men and 22 percent of women with depression reported that their depressive symptoms make it difficult for them to work, accomplish tasks at home, or get along with other people (Pratt & Brody, 2008).
Effects of depression in adolescents are similar to those in adults; however, Siu (2016) noted depression has a negative effect on developmental trajectories in children and adolescents younger than 18 years old. Also, major depressive disorder in the adolescent population is especially problematic because it is linked with higher possibility of suicide attempt, death by suicide, and recurrence of the disorder in young adulthood.
Evidence strongly recommends screening for depression in adolescent and adult patients. Specifically, the USPSTF found convincing evidence that screening in primary care settings improves accurate identification of adolescent and adult patients with depression (Siu, 2016). Yet Borner et al. (2010) cite evidence that physicians are identifying and treating depression among adolescents even less than among adults, and that more than "70 percent of children and adolescents suffering from serious mood disorders go unrecognized or inadequately treated" (Borner, 2010, p. 948). Additionally, according to the 2016 USPSTF guideline for screening for depression in children and adolescents, only 36 to 44 percent of children and adolescents with depression receive treatment, further evidence that the majority of depressed children and adolescents go untreated. Although primary care providers (PCPs) are the first line of defense in detecting depression, studies show that PCPs fail to identify up to 50 percent of depressed patients, due to both lack of time and a lack of brief, sensitive, and easy-to administer psychiatric screening tools (Borner, 2010).
Finally, according to the 2016 USPSTF guideline for screening depression among adults, the United States spent about $22.8 billion on depression treatment in 2009, and an additional estimated $23 billion on lost productivity (Siu, 2016). This substantial economic burden warrants regular screening for depression, as screening is the first step in identifying those at risk for developing major depressive disorder and closing the performance gap.

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 (4b) under Usability and Use.*
Provider-level performance scores using both claims and registry data suggest that there are still gaps in care and opportunities for improvement.
Average Performance Rates by Year (PQRS – all reporting methods)*:
2011–82.6% (0.6% of eligible professionals reporting)
2012–65.2% (0.4% of eligible professionals reporting)
2013–71.0% (1.3% of eligible professionals reporting)
2014–52.4% (7.5% of eligible professionals reporting)
*From the 2014 PQRS Reporting Experience Report and Appendix

Claims submitted 1/1/2015 through 12/31/2015
Number of Providers 26,169
Number of cases reported with valid denominator criteria: 3,002,169
Average Unweighted Score 63.8%
Average Weighted Score 36.5%
Standard Deviation 45.9%
Min 0%
Max 100%
Interquartile range 100%
10th percentile 0%
20th percentile 0%
30th percentile 5.9%
40th percentile 85.7%
50th percentile 100%
60th percentile 100%
70th percentile 100%

80th percentile 100%
90th percentile 100%

Please note: The unweighted average measure is the aggregated score for entire population. The weighted average is the average provider-level score, which is weighted by the number of patients in the denominator of each provider's score. All other statistics are based on weighted provider-level scores.

Registry submitted 1/1/2015 through 12/31/2015
Number of Providers 7,027
Number of cases reported with valid denominator criteria 989,092
Average Unweighted Score 50.7%
Average Weighted Score 28.9%
Min 0%
Max 100%
Interquartile range 99.7%
10th percentile 0%
20th percentile 0.3%
30th percentile 2.2%
40th percentile 17.8%
50th percentile 50.8%
60th percentile 85.7%
70th percentile 100%
80th percentile 100%
90th percentile 100%

Please note: The unweighted average measure is the aggregated score for entire population. The weighted average is the average provider-level score, which is weighted by the number of patients in the denominator of each provider's score. All other statistics are based on weighted provider-level scores.

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 (4b) under Usability and Use.*
Below are aggregate performance rates by patients' age, race, and sex using 3,002,169 Medicare claims from calendar year 2015. These results represent only those providers who voluntarily reported this measure and may not be generalizable to the population of all eligible providers.

Age Groups
18–64: 35.7%
65+: 36.7%
(X2 = 207.5; df: 1; N: 3,002,107; p < 0.0001)
We excluded age category 12-17 due to small sample size

Race
Asian: 58.4%
Black: 26.8%

Hispanic: 43.2%
Native American: 73.9%
White: 37.1%
Other: 50.0%
Unknown: 38.9%
(X2 = 31,993.1; df: 6; N: 3,002,169; p < 0.0001)

Sex
Female: 37.6%
Male: 34.8%
(X2: 2,575.2; df: 1; N: 3,002,169; p < 0.0001)

Because Medicare claims do not provide data on patients' insurance status, socioeconomic status, and/or disability status, we were unable to determine the presence of disparities in performance based on these factors.

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

Race/ethnicity: Literature indicates that depression rates are higher in non-Latino black people than in their non-Latino white counterparts (Pratt & Brody, 2008). Clinical practice guidelines also indicate that minority racial and cultural groups in the United States are less likely to receive treatment for depression than white Americans (Trangle et al., 2016). Data collected from electronic health records of approximately 65,079 adult primary care patients from 2010 to 2012 showed that (1) individuals from minority groups are less likely to undergo screening for mental disorders, such as depression screening; (2) minority groups have less access to mental health care and receive less than adequate health care compared to non-Latino whites, and (3) women from racial/ethnic minority groups are less likely than white women to have access to mental health care (Hahm et al., 2015). Medicare beneficiary survey data analyzed by Akincigil et al. showed that about 6.4 percent of white Americans, 4.2 percent of black Americans, and 7.2 percent of Latino Americans had a diagnosis of depression. Among those diagnosed, 73 percent of whites received treatment (either with antidepressants, psychotherapy, or both); 60 percent of blacks received treatment; and 63.4 percent of Latinos received treatment (Akincigil et al., 2012). These findings are consistent with other studies that show depression is under-recognized and undertreated among adult minorities. According to Davis et al. (2011, p.1282), "Recent data suggest that the proportion of depressed adults who seek treatment is significantly lower among African Americans (53%) than among Caucasians (67%)."
Age: Literature indicates that depression rates are highest among adults ages 40 to 59 (Pratt & Brody, 2008).
Gender: Literature indicates that depression is more common in women than in men (Pratt & Brody, 2008). Studies showed that men were less likely than women to receive screening for mental health problems, such as depression (Hahm et al., 2015). Among Latino and Asian Americans, women were more likely than men to receive screening for depression and visit a health care provider for depression care after depression was detected. Asian and black Americans, particularly black women, were less likely to receive screening for depression and less likely to receive any depression care than their white and Latino counterparts (Hahm et al., 2015).
Socioeconomic status: People with incomes below the federal poverty line and in the 18-39 and 40-59 age brackets experience higher depression rates than those with higher incomes, although this disparity is not observable in other age categories (Pratt & Brody, 2008).
We did not find any literature related to disparities associated with insurance status or disability.

Akincigil, A., Olfson, M., Siegel, M., Zurlo, K. A., Walkup, J. T., & Crystal, S. (2012). Racial and ethnic disparities in depression care in community-dwelling elderly in the United States. American Journal of Public Health, 102, 2, 319-328.
Davis, T. D., Deen, T., Bryant-Bedell, K., Tate, V., & Fortney, J. (2011). Does minority racial-ethnic status moderate outcomes of collaborative care for depression? Psychiatric Services, 62, 1282-1288.
Hahm, H. C., Cook, B. L., Ault-Brutus, A., & Alegri´a, M. (2015). Intersection of race-ethnicity and gender in depression care: Screening, access, and minimally adequate treatment. Psychiatric Services, 66, 258-264.
Pratt, L. A., & Brody, D. J. (2008). Depression in the United States household population, 2005-2006 (NCHS Data Brief No. 7). Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Trangle, M., Gursky, J., Haight, R., Hardwig, J., Hinnenkamp, T., Kessler, D., Myszkowski, M. (2016, March). Adult depression in primary care. Bloomington, MN: Institute for Clinical Systems Improvement. Retrieved from https://www.icsi.org/guidelines__more/catalog_guidelines_and_more/catalog_guidelines/catalog_behavioral_health_guidelines/depression/

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children, 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.)*
https://www.cms.gov/apps/ama/license.asp?file=/PQRS/downloads/2016_PQRS_IndMeasuresSpecs_ClaimsRegistry_022316.zip

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_0418_Coding_Table_S2b._3148_PQRS_134.xlsx

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.
We made minor denominator coding updates for the 2015 program year; see release note or code table in S.2b for specific details. We did not make any updates for the 2016 program year. We are making updates for the 2017 program year, which will be published after this submission. Changes for program year 2017 include: addition of Patient Health Questionnaire (PHQ-9) and Pediatric Symptom Checklist (PSC-17) to the Definition section of the specification; addition of examples of depression screening tools to clarify available standardized options for provider use, including depression screening tools for adolescents; CPT coding changes per expert panel recommendations, for example, deletion of one CPT code (90839) for the 2017 program year; changed

term clinical depression to depression because the word clinical could reduce the sensitivity of screening; and incorporated new literature into rationale.

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 screened for clinical depression on the date of the encounter using an age appropriate standardized tool AND, if positive, a follow-up plan is documented on the date of the positive screen

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).
Numerator Quality-Data Coding Options for Reporting Claims and Registry Satisfactorily:
G8431: Screening for clinical depression is documented as being positive AND a follow-up plan is documented
OR
G8510 Screening for clinical depression is documented as negative, a follow-up plan is not required

G8432 Clinical depression screening not documented, reason not given
OR
G8511 Screening for clinical depression documented as positive, follow-up plan not documented, reason not given

Definitions in relation to the Numerator include:
Screening – Completion of a clinical or diagnostic tool used to identify people at risk of developing or having a certain disease or condition, even in the absence of symptoms.

Standardized Depression Screening Tool – A normalized and validated depression screening tool developed for the patient population in which it is being utilized. The name of the age appropriate standardized depression screening tool utilized must be documented in the medical record.
Examples of depression screening tools include but are not limited to:

Adolescent Screening Tools (12-17 years) Patient Health Questionnaire for Adolescents (PHQ-A), Beck Depression Inventory-Primary Care Version(BDI-PC), Mood Feeling Questionnaire (MFQ), Center for Epidemiologic Studies Depression Scale (CES-D), and PRIME MD-PHQ2

Adult Screening Tools (18 years and older)
Patient Health Questionnaire (PHQ9), Beck Depression Inventory (BDI or BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Depression Scale (DEPS), Duke Anxiety-Depression Scale (DADS), Geriatric Depression Scale (GDS), Cornell Scale Screening, and PRIME MD-PHQ2

Follow-Up Plan- Documented follow-up for a positive depression screening must include one or more of the following:
•Additional evaluation for depression
•Suicide Risk Assessment
•Referral to a practitioner who is qualified to diagnose and treat depression
•Pharmacological interventions
•Other interventions or follow-up for the diagnosis or treatment of depression

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 12 years and older

S.7. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The denominator is defined by the patient´s age, encounter date, denominator CPT or HCPCS codes.
Patients aged > = 12 years on date of encounter AND
90791, 90792, 90832, 90834, 90837, 90839, 92625, 96116, 96118, 96150, 96151, 97003, 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, G0101, G0402, G0438, G0439, G0444

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Not Eligible – A patient is not eligible if one or more of the following conditions are documented:
•Patient refuses to participate
•Patient is in an urgent or emergent situation where time is of the essence and to delay treatment would jeopardize the patient's health status
•Situations where the patient's functional capacity or motivation to improve may impact the accuracy of results of standardized depression assessment tools. For example: certain court appointed cases or cases of delirium
•Patient has an active diagnosis of Depression
•Patient has a diagnosed Bipolar Disorder

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Denominator Exclusions are identified with the following provider reported HCPCS numerator clinical quality codes:

G8433 Screening for clinical depression not documented, documentation stating the patient is not eligible
OR
G8940 Screening for clinical depression documented as positive, a follow-up plan not documented, documentation stating the patient is not eligible.

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.)*
No stratification.

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.)*
PERFORMANCE CALCULATION – Claims and Registry

To calculate provider performance, complete a fraction with the following measure components: Numerator (A), Performance Denominator (PD) and Denominator Exclusions (B).

Numerator (A): Number of patients meeting numerator criteria

Performance Denominator (PD): Number of patients meeting criteria for denominator inclusion

Denominator Exclusions (B): Number of patients with valid exclusions

1) identify the patients who meet the eligibility criteria for the denominator (PD) which includes patients who are 12 years and older with appropriate encounters as defined by encounter codes or encounter value set during the reporting period.

2) identify which of those patients meet the numerator criteria (A)

3) for those patients who do not meet the numerator criteria, determine whether an appropriate exclusion applies (B) and subtract those patients from the denominator with the following calculation: Numerator (A)/[Performance Denominator (PD) - Denominator Exclusions (B)]

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

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.)*
IF a PRO-PM, 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 (Only), Registry

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
No specific data source/data collection instrument.

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)*
Clinician Office/Clinic
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 a composite.

2. Validity – See attached Measure Testing Submission Form
Testing_form_NQF_0418_3148__PQRS_134.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. (Do not remove prior testing information – include date of new information in red.)

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. (Do not remove prior testing information – include date of new information in red.)

No

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

No - This measure is not risk-adjusted

Measure Number (*if previously endorsed*)**:** 0418/3148
Measure Title: Preventive Care and Screening: Screening for Depression and Follow-Up Plan
Date of Submission: 12/9/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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

Part B Medicare claims data

Physician Quality Reporting System (PQRS) administrative data from registries

Previous Submission:
Claim Type: Claim Carrier (B)
Criteria: Any HCPCS Line code in the following string: G8431, G8510, G8433, G8432, G8511
Additional fields requested to the standard layout: LINE_PRCSG_IND (included in the detail file), beneficiary name, beneficiary DOB, beneficiary DOD, beneficiary gender, beneficiary HIC, and beneficiary race
NPIs who had fewer than ten claims were removed from the dataset. A simple random sample of records for approximately 150 NPIs was drawn. From those 150 NPIs, a random sample of approximately 600 claims was identified. The records were then stratified by the business location address listed in the NPI registry so that the maximum number of records from each business location was limited to 10 records. This limitation was set so that the providers would not see this task as too burdensome and would be more likely to send in their records.

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

We tested the measure using Part B Medicare claims data for encounters from 1/1/2015 to 12/31/2015.

We also tested the measure using the PQRS administrative registry data aggregated at the provider level for encounters from 1/1/2015 to 12/31/2015.

Previous Submission:
Time period: 1/1/2012 – 3/31/2012

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

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

Previous Submission:
N/A

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

We used Part B Medicare claims data reported by 26,169 providers, with an average of 115 patients in the denominator per provider. We used all claims with the following quality data codes (QDCs): G8431, G8432, G8433, G8510, G8511, G8940.

We used PQRS administrative registry data reported by 7,027 providers at 1,727 practices, with an average of 141 patients in the denominator per provider and an average of 550 patients in the denominator per practice.

Previous Submission:
Data Sample Response Rates:
Number of provider requested / returned / reviewed: 155/79/77
Provider response rate: 51.0%

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*)

Total number of patients with valid denominator criteria:
Claims data: 3,002,169
Registry data: 989,092

Descriptive characteristics of patients – claims data:
Sex
Female: 1,759,168 (58.6%)
Male: 1,243,001 (41.4%)

Race
Asian: 30,234 (1.0%)
Black: 363,378 (12.1%)
Hispanic: 57,189 (1.9%)
Native American: 10,992 (0.4%)
White: 2,472,318 (82.4%)
Other: 36,502 (1.2%)
Unknown: 31,556 (1.1%)

Age
12-17: 62 (0%)
18-64: 589,536 (19.6%)
65+: 2,412,571 (80.4%)

PQRS administrative data from registries did not include descriptive patient data.

Previous Submission:
Data Sample Response Rates:
Number of records requested / returned / reviewed: 641/294/275
Provider response rate 45.9%

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.

We used both sources (claims and registry) to assess providers' measure performance and provider-level reliability. We also used registry data to assess practice-level measure performance and reliability. We examined demographics and disparities in claims data but not registry data because registry data do not include patient characteristics such as age, sex, or race.

Previous Submission:
N/A

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

We aggregated performance scores in claims data by race, sex, and age to look for disparities. Claims data do not include information about income or other sociodemographic information. Registry data are aggregated at the provider-level and do not include sociodemographic variables.

Previous Submission:
N/A

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*)

Previous Submission:
[Critical data elements]

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

We calculated reliability using a widely accepted method that is outlined in J.L. Adams' (2009) technical report titled "The Reliability of Provider Profiling: A Tutorial." In this context, reliability represents a measure's ability to confidently

distinguish the performance of one physician from another. As discussed in the report, "Conceptually, [this method assesses] the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error." In this method, reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or variation across patients within providers), whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities. Although, there is not a clear cut-off for minimum reliability level, values above 0.7 are considered sufficient to see differences between physicians (or practices) and the mean, and values above 0.9 are considered sufficient to see differences between individual physicians or practices.

Adams, J.L. (2009). *The reliability of provider profiling: A tutorial* (TR-653-NCQA). Santa Monica, CA: RAND Corporation. Retrieved November 14, 2016, from http://www.rand.org/pubs/technical_reports/TR653.html

Previous Submission:

Crude agreement rates were calculated along with prevalence adjusted kappa (PAK), Cohen´s kappa values and corresponding confidence intervals. Cohen´s kappa represents chance-corrected proportional agreement. High prevalence of responses in a small number of cells is known to produce unexpected results known as the "kappa paradox" When the prevalence of a rating in the population is very high or low, which was noted in the testing of this measure, the value of kappa may indicate poor reliability even with a high observed proportion of agreement. In such cases, as with this measure, PAK is shown to provide an additional interpretation of agreement when the prevalence of responses is concentrated in a small number of cells.

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*)

Performance measure score reliability:

Data source	Number of providers/ practices	Between-provider variance	Reliability mean	Reliability median	Reliability standard deviation	Reliability min/max
Claims	26,169	.21	.99	1.0	.03	.62 - 1.0
Registry – provider level	7,027	.19	.99	1.0	.04	.60 - 1.0
Registry – practice level	1,797	.18	.99	1.0	.05	.59 - 1.0

Previous Submission:

Overall Reliability: Claims vs. Independent Review:
Numerator: 79.2% agreement, PAK=.60 (CI .51 -.69), Kappa=.38(CI .25 - .50)
Denominator Exclusions: 93.0% agreement, PAK=.86 (CI .80 - .92), Kappa .64 (CI .49 - .79)
Valid Denominator Criteria: 100.0%

Of the 275 total cases reviewed, 240 cases were not reported as exclusions and were reviewed for numerator agreement as compared with the code reported on the claim.
In 191 of 240 (79.0%) cases, reviewers agreed with whether or not the case met the numerator criteria based on the code submitted with the claim. A large proportion of those cases fell into the "met numerator criteria" category. Of the 191 cases where agreement was present, 165 cases(86.4%) agreed the case met the numerator criteria while 26 cases(13.6%)

agreed the case did not meet the numerator criteria. This prevalence should be considered in the interpretation of kappa scores.
Agreement with claims on denominator exclusions was high, with 93% agreement and prevalence adjusted kappa of .86 (95% CI .80 – .92).

Inter-Rater Reliability: Quality Insights of Pennsylvania Internal RN Reviewer (QIP) vs. Independent External RN Reviewer (ALPS):
Numerator: 89.7% agreement, PAK=.80 (CI .70 -.89), Kappa=.75 (CI .64 - .86)
Denominator Exclusions: 66.5% agreement, PAK=.39 (CI .30 -.48), Kappa .18 (CI .09 -.27)
Valid Denominator Criteria: 100%

All records without valid denominator criteria were removed prior to reliability assessment. Denominator agreement was 100%.

Reporting of this measure demonstrates high numerator reliability. Disagreements between ALPS and claims were identified when the ALPS abstractors was unable to find the documentation of the Age Appropriate Standardized Screening Tool used to perform the depression screening. The 2012 PQRS specification does not directly instruct the provider to document the tool used. Updating the measure specifications to add specific guidance to the provider to include documentation of the tool used in the patient's medical record could help to improve the reliability of this measure. Changes are in the process of being modified for the measure specification which will be finalized for the 2014 reporting year.

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?*)

With average reliability scores of 0.99 for both claims and registry reported data, this measure demonstrates a high level of reliability to detect real difference in performance scores.

Previous Submission:
Results of inter-rater reliability for two independent reviewers reflect differences in interpreting the exclusion criteria. Follow-up debriefings determined the main reason for mismatch was the use of different data sources to determine eligibility for exclusion. Upon review of the 2012 PQRS specification, the exclusion criterion was found to lack clarity which would have minimized the variability of interpretation when identifying exclusions. Additionally, while re-tooling this measure for an EHR specification in early 2012, this lack of specificity of the exclusion criterion was identified by Quality Insights of Pennsylvania (QIP) e-specification team. As a result, the 2013 PQRS Claims and Registry specification was updated and the exclusion criteria were changed to add more clarity. Additionally, the 2014 EHR specifications were re-tooled with updated exclusion language replacing the previous exclusions "Patient was referred with a diagnosis of depression" and "Patient has been participating in on-going treatment with screening of clinical depression in a preceding reporting period" with "Patient has an active diagnosis of depression or bipolar disease." Guidance will be added to the all specifications to define "active diagnosis" to reduce variability in this definition.

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of <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*)

Previous Submission:
[Face validity]

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

We surveyed 12 clinicians eligible to report this measure—none of whom advised on measure development—to rate face validity. We provided measure specifications and asked them to rate their agreement with the following statement: "The performance scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."

The rating scale offered five options: 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; 5 = strongly agree.

Previous Submission:
Quality Insights conducts an annual environmental scan to evaluate the most current research and evidence-based guidelines. A Technical Expert Panel (TEP), composed of subject matter specialists and experts with technical measure expertise, evaluates the results of the scan and provides recommendations based on the scientific merits of the evidence using the Strength of Recommendation Taxonomy (SORT). The TEP also reviews and establishes the measure's capability to capture what it is designed to capture using a consensus process.

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

1 – Strongly disagree – 0 votes
2 – Disagree – 3 votes
3 – Neither agree nor disagree – 0 votes
4 – Agree – 6 votes
5 – Strongly agree – 3 votes

Previous Submission:
N/A

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

Nine of 12 experts (75 percent) agreed or strongly agreed that the measure accurately reflects quality. Experts who disagreed raised concerns related to patient compliance, documentation burden, and a preference that the measure specify one screening tool for adolescents, rather than several tools from which providers can choose.

Previous Submission:

Face validity is established by subject matter specialists and experts who determine that the measure represents the process of interest.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

Previous Submission:
N/A

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We assessed the frequency of exclusions using claims and registry data from 2015.

Previous Submission:
N/A

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

In 2015 Medicare claims, 3.6 percent of eligible encounters qualified as exclusions. In 2015 registry data, 4.9 percent of eligible encounters were reported as exclusions.

Previous Submission:
N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The low rates suggest that the exclusions will not unduly distort measure performance. Although they may not dramatically change measure performance, exclusions allow for provider discretion in determining whether to screen patients for depression and provide follow-up interventions, improving the measure's face validity.

Previous Submission:
N/A

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

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

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

2b4.1.1 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

2b4.2. 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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

N/A

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

N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

N/A

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

N/A

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

N/A

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

N/A

2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

N/A

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

N/A

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

We used claims and registry data from calendar year 2015 to calculate measure performance scores and assess the distribution of performance using statistical measures of central tendency (mean and median), variation (standard deviation), and spread (interquartile range and rates by percentile).

Using claims data, we calculated chi-square statistics to test for significant differences between expected and observed aggregate performance scores based on patients' race, sex, and age. Registry data do not include descriptive patient data.

Previous Submission:

Dates of service from 1/1/2012 to 3/31/2012
Total Claims Submitted with any G code (G8431, G8510, G8433, G8432, G8511): 10,004
Valid Denominator Criteria: 7709 (77.1% of total)
Performance Exclusion: 1126 (14.6% of valid submissions)

Total tested claims sampled and reviewed: 275 records from 77 providers
Valid denominator criteria: 275/275 (100.0% of total)
Sample Performance Exclusion (claims based): 35 (12.7% of valid)
Aggregate measure performance rate for sample (claims based): 216/275 (78.5%)

Aggregate and provider (NPI) performance rates were calculated from Part B claims with dates of service from 1/1/2012 through 3/3/2012. Data from the testing sample were analyzed at the provider level. Performance rates are derived from G codes submitted for the Physician Quality Reporting System (formerly PQRI). Code submissions are voluntary and providers who report may not be representative of all eligible professionals. Performance rates cannot be generalized to the population.

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

Distribution of performance scores

Registry data (Provider-level)

- N: 7,027
- Mean: 50.7%
- STD: 44.3%
- Interquartile range: 99.7%
- 10th percentile: 0.0%
- 20th percentile: 0.0%
- 30th percentile:2.2%
- 40th percentile: 17.8%
- Median: 50.8%
- 60th percentile: 85.7%
- 70th percentile: 100.0%
- 80th percentile: 100.0%
- 90th percentile: 100.0%

Registry data (Practice-level)

- N: 1,797
- Mean: 55.2%
- STD: 41.8%
- Interquartile range: 95.0%
- 10th percentile: 0.0%
- 20th percentile: 1.2%
- 30th percentile: 12.1%
- 40th percentile: 39.9%
- Median: 63.4%
- 60th percentile: 85.7%
- 70th percentile: 99.2%
- 80th percentile: 100.0%
- 90th percentile: 100.0%

Claims data (Provider-level)

- N: 26,169
- Mean: 63.8%
- STD: 45.9%
- Interquartile range: 100.0%
- 10th percentile: 0.0%
- 20th percentile: 0.0%
- 30th percentile: 5.9%
- 40th percentile: 85.7%
- Median: 100.0%
- 60th percentile: 100.0%
- 70th percentile: 100.0%
- 80th percentile: 100.0%
- 90th percentile: 100.0%

Performance results by population groups, claims data:

Age groups
18–64: 35.7%
65+: 36.7%
(X^2 = 207.5; *df*: 1; *N*: 3,002,107; $p < 0.0001$)
(Age category 12–17 was excluded due to small sample size)

Race
Asian: 58.4%
Black: 26.8%
Hispanic: 43.2%
Native American: 73.9%
White: 37.1%
Other: 50.0%
Unknown: 38.9%
(X^2 = 31,993; *df*: 6; *N*: 3,002,169; $p < 0.0001$)

Sex
Female: 37.6%
Male: 34.8%
(X^2: 2,575.2; *df*: 1; *N*: 3,002,169; $p < 0.0001$)

Previous Submission:
Performance rates were calculated at the provider level for claims from the time period 1/1/2012 to 3/31/2012.
Aggregate measure performance rate: 5463/6583 (83.0%)
Distribution of provider scores (by NPI): N=459, Mean = 84.3%, Median=100.0%, SD=.339 Range=100
10th percentile: 0%, 25th percentile: 100.0%; 75th percentile 100.0%

Testing sample
Measure performance rate (claims based): 216/275 (78.5%)

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

Reported performance rates indicate a wide degree of variation in both claims and registry data, and many clinicians have the potential to improve the rates of depression screening and follow-up. The differences in performance rates by age and sex were statistically significant but small in magnitude and therefore of limited clinical significance. Among racial groups, clinically significant differences are apparent, and quality improvement efforts should attempt to address these disparities. However, we did not stratify the measure based on race because: (1) many other process measures show similar racial disparities and (2) stratifying the measure would significantly complicate implementation, reporting, and interpretation. Providers report this measure voluntarily, and reported performance rates may not represent the total eligible population.

Previous Submission:
N/A

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

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

Claims/registry and electronic clinical quality measure (eCQM) specifications are aligned across reporting methods. As directed by NQF, eCQM testing data are submitted separately as NQF 3132.

Previous Submission:
N/A

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Entities report the measure using a single data source. We did not compare performance rates between the claims/registry measure and the eCQM because the eCQM is submitted separately. However, we designed the specifications for all the data sources to maximize alignment and consistency.

Previous Submission:
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

Previous Submission:
N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A
Previous Submission:
N/A

_____
2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

There were no missing data in our analysis because our testing relied on all of the cases reported for this measure via claims and registries for the PQRS program from 1/1/2015 to 12/31/2015. In claims data, if an encounter record included a relevant CPT encounter code and QDC, we assumed it was eligible for the measure, and we used QDCs to group encounters into performance categories (met performance, failed performance, excluded). If these data were not available, we did not include the encounter in our analysis. Our registry data source included provider-level results for all providers who participated in registry reporting to the PQRS program. These data showed the number of patients per provider who met performance, failed performance, and were excluded.

Previous Submission:
N/A

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

There were no missing data in our claims or registry analyses. As noted above, claims lacking QDCs were assumed to be not relevant and excluded from our analysis.

Previous Submission:
N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling

of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

We used data from all eligible professionals who reported the measure using claims or registry data, which meant there were no systematic missing data. Because reporting is voluntary, the reporting population is not necessarily representative of the total eligible population and results are not generalizable to the overall eligible population. According to the appendix tables in the 2014 PQRS Reporting Experience, 7.5 percent of eligible professionals reported the measure in 2014 (Centers for Medicare & Medicaid Services, 2016).

Centers for Medicare & Medicaid Services. (2016, April). *2014 PQRS reporting experience, including trends (2007-2015).* Retrieved November 14, 2016, from https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

Previous Submission:
N/A

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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
We did not encounter any difficulties related to data availability. Based on our experience working with providers who report this measure for CMS quality reporting programs, the measure is feasible and its reporting is facilitated by the use of Quality Data Codes in claims and registry data that identify encounters that meet or fail to meet performance, or are ineligible or excluded from performance. This measure is not a PRO-PM.

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)*.
None

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 Physician Quality-Reporting System http://www.cms.gov/PQRS

4a.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

The Physician Quality Reporting System (PQRS), sponsored by Centers for Medicare & Medicaid Services, is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). To be eligible for an incentive payment, EPs must satisfactorily report data on quality measures for covered Physician Fee Schedule services furnished to Medicare Part B Fee-for-Service beneficiaries. More information about PQRS is available at http://www.cms.gov/PQRS. According to the 2014 PQRS Reporting Experience, in 2014, this measure was one of six program measures in which more than 500,000 professionals were eligible to report, yet only 7.5 percent of those eligible actually reported. EP performance scores that rely on registry reporting are posted on Physician Compare.

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

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

4b. 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.
Average PQRS reporting rates from 2011 to 2014 reflect reporting by all participating providers, including those who reported the measure using EHR, claims, and registry data. EPs submit performance data voluntarily, and results may not be representative of all EPs. We do not have access to data on historical trends in performance specific to claims and registry reporting, nor on performance rates by geographic area.

The average performance rate has fluctuated substantially over the past four years, decreasing from 82.6 percent in 2011 to 52.4 percent in 2014. However, the number of EPs reporting the measure has increased significantly over this time frame, from just 0.6 percent of EPs in 2011 to 7.5 percent in 2014. This makes it difficult to assess trends over time, as the EPs who recently began

reporting the measure may have lower performance rates than those who have been reporting it for a longer period. Although the reporting increased each year, a substantial number of EPs are still not reporting the measure, and the average performance rate illustrates that there is still a gap in care.

4c. Unintended Consequences

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

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

We have not identified any unintended consequences in our recent testing, or in the measure's implementation.

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

We have not identified any unexpected benefits in our recent testing, or in the measure's implementation.

4d1.1. **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

4d1.2. **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

4d2.1. **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

4d2.2. **Summarize the feedback obtained from those being measured.**

N/A

4d2.3. **Summarize the feedback obtained from other users**

N/A

4d.3. **Describe how the feedback described in 4d.2 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

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)**
0518 : Depression Assessment Conducted

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
There are no competing measures. Multiple related measures have lost their NQF endorsement, including:
•Percent of Residents Who Have Depressive Symptoms (Long-Stay) – Centers for Medicare & Medicaid Services (formerly NQF #0690)
•Depression Screening by 13 Years of Age – National Committee for Quality Assurance (formerly NQF #1394)
•Maternal Depression Screening – National Committee for Quality Assurance (formerly NQF #1401)
•Depression Screening by 18 Years of Age – National Committee for Quality Assurance (formerly NQF #1515)

We also identified the following measures in the National Quality Measures Clearinghouse that do not have NQF endorsement:
•Adult depression in primary care: percentage of perinatal patients with documentation of screening for major depression or persistent depressive disorder using either PHQ-2 or PHQ-9 (Institute for Clinical Systems Improvement [ICSI])
•Adult depression in primary care: percentage of patients with cardiovascular disease with documentation of screening for major depression or persistent depressive disorder using either PHQ-2 or PHQ-9 (ICSI)
•Adult depression in primary care: percentage of patients who had a stroke with documentation of screening for major depression or persistent depressive disorder using either PHQ-2 or PHQ-9 (ICSI)
•Pediatric preventive care: percentage of pediatric patients aged 12 to 17 years who have a documented mental health and/or depression screening using one of the specified validated tools at a well-child visit during the measurement period (Minnesota Community Measurement)

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 only related NQF endorsed measure identified is 0518: Depression Assessment Conducted. Measure 0518 is an episode-based measure and reported based on OASIS data specific to home health agencies. It is similar to 0418, as it assesses depression using a standardized tool, but it differs in two key ways: First, target population: the denominator incorporates only adults aged 18 years and older and includes the number of home health episodes of care ending during the reporting period. Second, measure focus: the measure focuses on home health care in which patients received screening for depression. It does not include any follow-up component. 0418 is a patient-based measure focused on patients 12 years and older and includes a follow-up plan for positive depression screening results. Both are process measures; however, data for 0518 are only reported electronically and 0418 data may be reported using claims, registry, and electronic sources. 0418 is more robust in that it includes a broader population and requires a follow-up plan of care.

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 that target the same measure focus and or 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.

Attachment Attachment: NQF_0418_Summary_Materials.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Sophia, Autrey, Sophia.Autrey@cms.hhs.gov
Co.3 **Measure Developer if different from Measure Steward:** Quality Insights of Pennsylvania
Co.4 **Point of Contact:** Anita, Somplasky, asomplasky@wvmi.org, 877-346-6180-7852

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.
Through a collaborative process, the expert workgroup annually reviews the measure specifications (description, numerator, denominator, definitions, and clinical recommendation); literature review findings; and feedback or questions about the measure during its implementation. When last convened in 2016, the expert workgroup included the following members:

Jean Carter, PhD
Psychology
Washington Psychological Center, P.C.

Paula Hartman-Stein, PhD
Clinical psychology
Center for Healthy Aging; clinical psychologist, founder

Bracken Babula, MD
Internal medicine
Department of Medicine; Thomas Jefferson University; associate quality officer

Alan Axelson, MD
Adolescent psychiatry
InterCare Psychiatric Services; medical director and chief

Justin Schreiber, DO, MPH
Psychiatry
Western Psychiatric Institute and Clinic; co-triple board chief

Gregory M. Martino, PhD
Clinical psychology
Independent practice, DuBois, Pennsylvania

Tracy Murphy, AuD
Audiology
North Shore Audio-Vestibular Lab

Virginia Clark, PhD
Psychology (adolescent)
Western Reserve Psychological Associates, Inc.; president

Donald Wilson, MD
Obstetrics/gynecology
Women's Care Florida; chief medical officer

Harold Manley, PharmD
Pharmacology
Dialysis Clinic, Incorporated; director of medication management and pharmacovigilence

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2008

Ad.3 **Month and Year of most recent revision:** 09, 2016

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?** 09, 2017

Ad.7 **Disclaimers:** This measure and specifications are provided "as is" without warranty of any kind. This measure does not represent a practice guideline.

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

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 3172
Corresponding Measures:
Measure Title: Continuity of Pharmacotherapy for Alcohol Use Disorder
Measure Steward: RAND Corporation
Brief Description of Measure: Percentage of adults 18-64 years of age with pharmacotherapy for alcohol use disorder (AUD) who have at least 180 days of treatment and a Proportion of Days Covered (PDC) of at least 0.8
Developer Rationale: In spite of its high prevalence and its substantial burden on patients, their families and society, alcohol use disorder (AUD) remains a severely undertreated condition. According to the 2014 National Survey on Drug Use and Health (NSDUH), 16.3 million Americans ages 18 years and older suffered from AUD (SAMHSA, 2014a), representing almost 7 percent of the adult population (SAMHSA, 2014b). But only 15.2 percent of patients, who reported that they needed alcohol treatment, actually received it (SAMHSA, 2014c).

Medication-assisted treatment (i.e., pharmacotherapy combined with counseling) is an evidence-based and effective treatment option for patients with AUD. It is supported by several national guidelines and a Consensus Panel of the National Institute on Alcohol Abuse and Alcoholism and the Substance Abuse and Mental Health Services Administration (SAMHSA, 2015), but substantially underused. A recent study that included 16,947 AUD patients treated in a network of community health centers, for example, found that only 3.2 percent of those patients received pharmacotherapy (Riekmann et al., 2016). Minority patients and those without health insurance were significantly less likely to receive pharmacotherapy.

For patients receiving pharmacotherapy, continuity of treatment is critical. Most experts believe that substance dependence, including AUD, should be considered and treated as a chronic illness (McLellan, 2000). This is particularly important for those patients who seek and receive treatment (McLellan, 2002), as they have self-identified to be at risk for a chronic relapsing course (Dawson, 1996; Institute of Medicine, 1998; McKay, 2005; Miller & Hester, 1986).

Overall, longer duration of AUD treatment is associated with better outcomes (Lemke & Moos, 2003; Moos et al., 1995; Oimette et al., 1998) and treatment adherence is essential, as medication non-adherence is associated with relapse to heavy and/or frequent drinking and higher health care utilization (Gueorguieva et al., 2013; Kranzler et al., 2008; Stout et al., 2014). Yet evidence suggests that persistence of pharmacotherapy is poor. Baser et al. looked at medication adherence for six months after initiation of treatment in a large database study including 15,502 patients treated with an FDA-approved AUD drug. They found that only six to eleven percent of patients on oral drugs and only 21 percent of patients with injectable naltrexone were sufficiently adherent (Baser et al., 2011).

Therefore, the proposed measure focuses on continuity of pharmacotherapy, defined as treatment duration of at least 180 days and sufficient adherence for the duration of treatment. The definition of adherence follows the established convention of having access to medication for at least 80 percent of treatment days.

Several important benefits related to quality improvement are envisioned with the implementation of this measure. First, the measure will help health plans and providers to identify individuals with AUD, who are non-adherent to or discontinue pharmacotherapy. As a result, this measure will encourage health plans and providers to develop communication and education tools and processes to improve treatment continuity in their patients with AUD. Improved treatment continuity is expected to result in lower rates of relapse, and less substance use-related morbidity and mortality. Adoption of this performance measure has the potential to improve quality of care for individuals with AUD and, therefore, advance quality of care by engaging patients as partners in their care, and promoting effective communication and coordination of care, priority areas identified in the National Quality Strategy.

CITATIONS

Baser O, Chalk M, Rawson R, Gastfriend DR. Alcohol dependence treatments: comprehensive healthcare costs, utilization outcomes, and pharmacotherapy persistence. Am J Manag Care [2011, 17 Suppl 8:S222-34].

Dawson DA. Correlates of past-year status among treated and untreated persons with former alcohol dependence: United States, 1992. Alcoholism, Clinical and Experimental Research. 1996;20(4):771-779.

Gueorguieva R, Wu R, Krystal JH, Donovan D, O'Malley SS. Temporal patterns of adherence to medications and behavioral treatment and their relationship to patient characteristics and treatment response. Addictive Behaviors. 2013;38:2119-21.

Institute of Medicine. Bridging the Gap Between Practice and Research: Forging Partnerships with Community-Based Drug and Alcohol Treatment. Washington, DC: The National Academies Press; 1998.

Kranzler HR, Stephenson JJ, Montejano L, Wang S, Gastfried DR. Persistence with oral naltrexone for alcohol treatment: implications for health-care utilization. Addiction. 2008;103:1801-1808.

Lemke S, Moos RH. (2003). Outcomes at 1 and 5 years for older patients with alcohol use disorders. J Subst Abuse Treat. 24(1):43-50.

McKay JR. Is there a case for extended interventions for alcohol and drug use disorders? Addiction. 2005;100(11):1594-1610.

McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. JAMA. 2000;284(13):1689-95.

McLellan AT. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. Addiction. 2002;97(3):249-252.

Miller WR, Hester RK. The Effectiveness of Alcoholism Treatment. In: Miller WR, Heather N, eds. Treating Addictive Behaviors: Processes of Change. Boston, MA: Springer US; 1986:121-174.

Moos RH, Pettit B, Gruber V. (1995). Longer episodes of community residential care reduce substance abuse patients´ readmission rates. J Stud Alcohol. 56(4):433-43.

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522.

Riekmann T, Muench J, McBurnie M, et al. Medication-assisted treatment for substance use disorders within a national community health center research network. Subst Abuse. 2016;37(4):625-634.

Stout RL, Braciszewski JM, Subbaraman MS, Kranzler HR, O'Malley SS, Falk D. What happens when people discontinue taking medications? Lessons from COMBINE. Addiction. 2014;109:2044-2052.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014a). 2014 National Survey on Drug Use and Health (NSDUH). Table 5.8A—Substance dependence or abuse in the past year among persons aged 18 or older, by demographic characteristics: Numbers in thousands, 2013 and 2014. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.htm#tab5-8a

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014b). 2014 National Survey on Drug Use and Health (NSDUH). Table 5.8B—Substance dependence or abuse in the past year among persons aged 18 or older, by demographic characteristics: Percentages, 2013 and 2014. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.htm#tab5-8b

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014c). 2014 National Survey on Drug Use and Health (NSDUH). Table 5.24A – Locations Received Alcohol Treatment in the Past Year among Persons Who Received Alcohol Treatment in the Past Year, by Age Group: Numbers in Thousands, 2013 and 2014. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.htm#tab5-24a

Substance Abuse and Mental Health Services Administration and National Institute on Alcohol Abuse and Alcoholism (SAMHSA). (2015). Medication for the Treatment of Alcohol Use Disorder: A Brief Guide. HHS Publication No. (SMA) 15-4907. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Numerator Statement: Individuals in the denominator who have at least 180 days of treatment and a PDC of at least 0.8 for AUD medications

S.6. **Denominator Statement:** Individuals 18-64 years of age who had a diagnosis of AUD and at least one claim for an AUD medication

Denominator Exclusions: There are no denominator exclusions.

Measure Type: Process
Data Source: Claims (Other), Pharmacy
Level of Analysis: Health Plan, Population : Regional and State

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report
1a. Evidence

1a. Evidence. The evidence requirements for a *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.

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 diagram of the relationship of this process of care (pharmacotherapy for Alcohol Use Disorder (AUD) proportion of days covered (PDC) > 0.8 for at least 180 days) to lower AUD relapse rates, which in turn leads to fewer adverse outcomes and decreased costs.
- Evidence provided by the developers to support the measure includes two recommendations from the VA/DoD 2015 Guideline on Management of Substance Use Disorders.
    - **Recommendation 5**: For patients with moderate-severe alcohol use disorder, we recommend offering one of the following medications (***recommendation grade: "strong for…")***:
        - Acamprosate (***evidence grade=moderate)***
        - Disulfiram (***not graded***; using the six Cochrane criteria for assessing risk of bias, 6 of 22 studies met none of the criteria, while the remainder met between 1 to 4 of the criteria)
        - Naltrexone- oral or extended release (***evidence grade=low to moderate***)
        - Topiramate (***not graded***, but relied on randomized, double-blind, placebo-controlled trials)
    - **Recommendation 6**: For patients with moderate-severe alcohol use disorder for whom first-line pharmacotherapy is contraindicated or ineffective, we suggest offering gabapentin. (***recommendation grade: "weak for…"; evidence not graded)***
    - For each of the named medications, in the "estimates of benefit and consistency across studies" section, the developers summarized the link between treatment and the likelihood of various types of relapse (e.g., return to drinking, return to heavy drinking, reduction in heavy drinking days). However, the developer did not summarize evidence showing that preventing relapse does, in fact, leads to fewer adverse health outcomes.
- The developer provided several citations to support their use of the 180-day minimum treatment period and the use of an 80% threshold for PDC.

Exception to evidence: N/A

Questions for the Committee:

- *Does reducing relapses in AUD patients lead to better health outcomes?*
- *How strong is the evidence that use of the various medications is associated with reductions in relapse?*

Guidance from the Evidence Algorithm

Process measure supported by systematic review and grading (Box 3) → QQC provided (Box 4) → evidence varied by medication type, but overall: Quantity: high; Quality: moderate; Consistency: moderate (Box 5b) →Moderate

The highest possible rating is HIGH.

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.

- Performance results for 2010-2015 (for two-year rolling periods) are calculated from commercial pharmacy and medical claims obtained from the Truven MarketScan® Commercial Database. Data

were limited to members with at least 6 months of continuous enrollment. The total number of episodes included in the analysis ranged from 11,737 in the 2010-2011 period to 26,803 in the 2014-2015 period.

Summary statistics of performance across states

Time Period	N	Mean	Min	Max	STD	IQR	P10	P25	Median	P75	P90
2010-2011	44	0.161	0.067	0.241	0.035	0.045	0.121	0.138	0.162	0.183	0.200
2011-2012	45	0.194	0.087	0.300	0.041	0.052	0.149	0.172	0.187	0.224	0.237
2012-2013	45	0.196	0.062	0.278	0.044	0.048	0.151	0.175	0.194	0.222	0.250
2013-2014	46	0.195	0.116	0.305	0.036	0.037	0.152	0.170	0.197	0.206	0.237
2014-2015	47	0.218	0.125	0.364	0.036	0.033	0.187	0.202	0.211	0.235	0.256

Summary statistics of performance across commercial health plans

Time Period	N	Mean	Min	Max	STD	IQR	P10	P25	Median	P75	P90
2010-2011	58	0.149	0	0.350	0.078	0.100	0.045	0.091	0.146	0.190	0.250
2011-2012	93	0.194	0	0.409	0.080	0.112	0.091	0.138	0.200	0.250	0.300
2012-2013	138	0.192	0	0.440	0.074	0.097	0.103	0.143	0.180	0.240	0.304
2013-2014	179	0.195	0	0.422	0.075	0.083	0.100	0.150	0.192	0.233	0.302
2014-2015	203	0.219	0.048	0.421	0.073	0.110	0.125	0.161	0.221	0.271	0.313

Disparities

- The developer provided overall performance results (not aggregated by state or health plan) by both age group and sex. Performance (i.e., percentage of AUD patients with PDC ≥80%) was higher for older versus younger age groups and for women compared to men. It is not clear whether or not the differences between the groups were statistically significant.

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 for other patient subpopulations?*

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 to Support Measure Focus

Comments:

**The developer defines "treatment" initially to include both counseling and pharmacotherapy. Yet this measure only looks at days patients receive pharmacotherapy. There is evidence to support the greater efficacy of counseling plus medication for increasing rates of recovery and contributing to better health outcomes. This definition of "treatment" is also inconsistent with the discussions around #004 which define treatment as psychotherapy alone, MAT alone, or both psychotherapy and MAT.

**There is ample evidence that too few patients receive AUD treatment and that only 3.2% receive pharmacotherapy. I do not understand why folks 65 and older were excluded.

**High, evidence is strong.

**The process does measure the length and compliance with medication treatment as a proxy linked to better outcomes supported by the research evidence. The evidence by drug treatment is inconsistent by medication. Yes the PRO is support by the stated rational

**The relationship between the intermediate outcomes of reduction in relapse and better health and between adherence to effective treatment and reduced relapse seem sound. Given that many of the relevant health outcomes are relatively rare or far in the future, it seems unreasonable to hold out for health outcomes in this subject matter

**I'd like to hear discussion from the committee about the size of the effects (which seem generally modest) and the value of breaking out drug therapy separately from talk therapy.

1b. Performance Gap

Comments:

**As the message developer points out in the rationale, there is a significant underutilization of MAT for AUD. This measure is focusing on the gap in continuity of care for individuals already prescribed MAT for AUD, but does not measure the gap in initial prescribing MAT for AUD. Nonetheless there is a performance gap from a continuity stand point as well.

**Performance Gap clearly exists

**Relatively low rates of performance by state (.161-.218) and commercial health plans (.149-.219). Level of analysis is at the state and health plan, not agency or provider.

**There is a need for a national performance measure and benchmarks. There are extremely limited metrics that are linked to a proxy for outcome like this one. With the difference in the prevalence in AUD in subpopulations, not clear if there has been an research that has tested this metric and identified similar benchmarks

**There certainly seem to be gaps in care as well as a trend toward improvement. There are also clearly variations around the fairly low general performance. There was some limited stratification (e.g., by sex and age) that seems to show some variation in performance.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

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

Data source(s): Administrative claims data, including pharmacy claims

Specifications:

- This measure is specified for the health plan and state/region levels of analysis in the behavioral health outpatient setting and clinician office/clinic settings. A higher score indicates better quality.
- Patients included in the measure denominator include those ages 18-64 with a diagnosis of AUD who are continuously enrolled in a commercial health plan and have at least one pharmacy claim for at least one AUD medication (Acamprosate; Disulfiram; Naltrexone (oral/ extended-release injectable); Topiramate; Gabapentin).
- Patients included in the measure numerator include those with 1) at least 180 days from Day 1 of the first AUD medication claim through the measure end date (date the supply from last claim is exhausted, death date, or Dec 31 of year 2 of the measurement period, whichever comes first) and 2) a PDC of ≥80% (covered days are summed across pharmacy claims based on fill date and days' supply).
- The measure appears to be limited to states or health plans with ≥20 patients in the denominator.
- Codes (ICD-9-CM, ICD-10-CM, NDC, HCPCS) and descriptions for the measure data elements are provided, either in the submission form itself or in the supplementary materials provided with the submission.
- No exclusions are defined for the measure (note that members who are not continuously for at least 6 months after the first AUD medication fill during the measurement period are not included in the measure).
- The developer suggests stratification of results according to age group, sex, state, and health plan. This stratification is not meant for risk-adjustment purposes.
- A detailed calculation algorithm is provided.
- This measure is not risk-adjusted.

Questions for the Committee:

- o *The care settings indicated in the "care setting" and "calculation algorithm" sections of the submission form are inconsistent. Should this measure apply to other settings of care, as indicated in the calculation algorithm?*
- o *Are all the data elements clearly defined? Are all appropriate codes included?*
- o *Is the logic or calculation algorithm clear?*
- o *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data used for testing included 2013-2014 commercial pharmacy and medical claims obtained from the Truven MarketScan® Commercial Database. Data were limited to members with at least 6 months of continuous enrollment (n=22,330). A total of 46 states and 179 "pseudo" health plans were included in the analysis (the Truven database includes information from self-insured employers but does not include a plan ID and therefore health plan membership was imputed based on industry type and MSA).
- Developers conducted a signal-to-noise analysis, which is an appropriate method for testing reliability. Specifically, they used the robust Prasad-Rao estimator to estimate the signal and the standard binomial distribution inference to estimate the noise. The calculations were based on the ***average denominator size*** for states (n=471) and health plans (n=33); see Table 1 for information regarding the distribution of patients across states and health plans. The analysis was limited to those health plans with at least 20 members who were eligible for the measure denominator.
- A signal-to-noise analysis quantifies the amount of variation in a performance measure that is due to true differences between providers (i.e., signal) as opposed to measurement error (i.e., noise). Results will vary based on the amount of variation between health plans (or states) and the number of patients treated by each health plan (or state). A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in health plan (or state) performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Results of reliability testing

- State: reliability=0.772; standard deviation=0.068
- Health plans: reliability=0.846; standard deviation=0.053

Questions for the Committee:

- o *Is the test sample adequate to generalize for widespread implementation?*
- o *Reliability was estimated based on the average patient count (471 for states and 33 for health plans). Typically, reliability is lower when the patient count is lower. Is it reasonable to assume that states and health plans will have enough patients eligible for the measure to ensure adequate reliability? If not, is there any data that would indicate how many states or health plans would not have sufficient numbers of patients?*
- o *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm
Precise specifications (Box 1) → Empirical reliability testing with measure as specified (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → Moderate certainty that measure results are reliable (Box 6b) → Moderate

The highest possible rating is HIGH.

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

2b. Validity

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence:
The evidence for gabapentin is weaker than that for the other medications included in the measure. The developer cites use of the 80% threshold as "established convention" but does not further summarize the literature supporting this decision.

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
- Face validity was assessed by a 10 clinicians with expertise in treating AUD. These individuals were asked to rate their agreement, on a 5-point scale, with the following statement: "*Performance scores resulting from the measure as defined can be used to distinguish good from poor quality*."
- According to NQF guidance, the 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 testing conducted by the developer conforms to NQF's requirements for face validity.

Validity testing results:
- Of the 10 clinicians surveyed, 2 strongly agreed, and 4 agreed, that results from the measure can be used to distinguish good from poor quality. The remaining 4 clinicians neither agreed nor disagreed.

Questions for the Committee:
- *Did the clinicians included in the face validity assessment have the appropriate expertise to judge the face validity of the measure?*

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

2b3-2b7. Threats to Validity

2b3. Exclusions:

- No exclusions are defined for the measure (note that members who are not continuously enrolled for at least 6 months after the first AUD medication fill during the measurement period are not included in the measure).

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Questions for the Committee:

- *Process measures generally are not risk adjusted. Do you agree with the developer's decision not to risk-adjust this measure?*

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

- To assess whether differences in performance between states and health plans are meaningful, developers constructed 95% confidence intervals around each state and health plan's performance rate and compared these to the overall state and health plan average performance rate, respectively. They considered the state or health plan rate to be statistically different from the average if the overall mean did not overlap the individual state/health plan confidence interval. The developer used the 2013-2014 Truven MarketScan® Commercial Database for the analysis.
- States
  - 3 of the 46 states (6.5 percent) had scores statistically significantly lower than the state-level mean
  - 3 of the 46 states (6.5 percent) had scores statistically significantly higher than the state-level mean
- Health Plans
  - 12 of 179 health plans (6.7 percent) had scores statistically significantly lower than the health plan-level mean
  - 5 of 179 health plans (2.8 percent) had scores statistically significantly higher than the health plan-level mean

Question for the Committee:

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

2b6. Comparability of data sources/methods: N/A

2b7. Missing Data

- The developers assessed the frequency of missing, zero, or negative values for the "days supply" data element. They report that 457 (0.7%) of the 62,309 individuals eligible for the measure had one or more pharmacy claims with an invalid value for this variable. They interpret this to mean that missing or invalid data for this data element would not substantially impact the measure results.

Guidance from the Validity Algorithm

Specifications somewhat consistent with evidence (Box 1) → potential threats to validity assessed (Box2) → empirical validity testing not conducted (Box 3) → face validity systematically assessed (Box 4) → results indicate only moderate agreement that the measure results can be used to distinguish quality (Box 5) →Moderate to Low

The highest possible rating is MODERATE.

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

Comments:

**I am unclear how the measure defines "outpatient" setting. Does this include primary care settings as well? If not it should. I question the inclusion of gabapentin and topiramate since they are not FDA approved for the treatment of an AUD.

**Unless the N is too small it would be advantageous to include other settings

**Specifications are clearly described.

**Codes were clearly identified. There was limited risk adjustment to sex and age. Not clear why the presence of mental health diagnosis was not considered as a risk-adjustment element. Dual diagnosis cases are more difficult to maintain in treatment adherence.

**In general the specifications seem reasonable. I would like to hear discussion among treatment experts about whether all the medications currently included should be included. I agree that the care settings should be clarified. The 20 patients in the denominator seems reasonable for generating interpretable data.

2a2. Reliability Testing

Comments:

**Unclear if numbers will be sufficient to ensure reliability especially in subpopulations and smaller settings

**Reliability testing is only based on beta binomial, with r=.772 (state), and r=.846 (health plan)

**The validity testing was by using face validity methodology. The face validity methodology followed appeared to meet NQR requirements.

**Seems reasonable.

2b1. Validity Specifications

Comments:

**I agree with the comments in the PA

**Data source: Truven MarketScan Commercial Data base. No Medicaid. Only self-insured employees? Examined performance by state and health plan; able to stratify by age group and gender. Assume no capacity to examine by race/ethnicity with this data source.

**Agree with the reviewer. the evidence on gabapentin is not as strong as was the evidence for Disulfiram.

**Seems reasonable.

2b2. Validity Testing

Comments:

**I'm concerned that only 6 out of 10 clinicians agreed that the measure can be used to distinguish good versus bad quality

**Validity is based solely on two items from a survey monkey survey of 10 clinicians. Selection of this panel is not described. Weird: Dr. Chung is a CAP, full time employee at DMH based at Harbor and works under Dr. Ken Wells to support community engagement component of his R01. Was this a convenience sample of clinicians that may be affiliated with RAND through subcontracts?

For face validity: 2=strongly agree, 4=agree, and 4=neither agree or disagree

For usability: wide spread: 4=strongly agree, 3=agree, 1=neither; 2=disagree

Meaningful difference is solely based on # the states and health plans that had an average performance rate outside the 95% CI. (3 (6.5%) of states, 12 (6.7%) and 5 (2.8%) of health plans.

**Given the research over several different drugs with similar conclusion about length of medication treatment and in other research that demonstrates that the length of case management adherence correlates to better outcome, this measure is a acceptable indicator of guideline compliance and maintaining individuals in treatment. Definition of treatment quality should include improvement in functional status. Not sure if maintaining in treatment through these studies also tested functional changes as an outcome.

**Could the developer provide additional information on why the experts who did not agree that the measure had face validity for distinguishing better from worse performance came to that conclusion?

2b3. Exclusions Analysis

2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures

2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**If numbers are sufficient it would be helpful to risk adjust this
**Not clear if the missing data claims were from a systems with monthly refill claims of mail order based claims where billing is every three months. This was not addressed.
**Seems likely to show meaningful differences

Criterion 3. Feasibility

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

- The required clinical data elements (e.g., diagnosis) are routinely generated and used during care delivery
- The required data elements are available in electronic sources (i.e., administrative medical and pharmacy claims)
- The developers calculated measure results using a publicly-available claims database. They did not identify any feasibility or implementation issues.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
**It's feasible
**Difficult to assess. Not in operational use. Data source does not include Medicaid. Not vetted.
**There was no mention of the difference in pharmacy claims generation based on whether scripts are filled retail or through warehouse mailing in a 3 month intervals. Pharmacy claims may not be available in EMR.
**Because this measure is based on claims, it seems feasible

Criterion 4: Usability and Use

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

Current uses of the measure:
Publicly reported? ☐ **Yes** ☒ **No**

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

- The developers suggest this measure is appropriate for use in the CMS Medicaid Adult Core Set. They plan to explore avenues for recommending this measure for use in this program (i.e., potentially through the annual rule-making process)

Accountability program details: N/A

Improvement results. New measure – none reported.

Unexpected findings (positive or negative) during implementation. None reported.

Potential harms: None reported.

Vetting of the measure: None reported.

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 *Would the benefits of the measure outweigh any potential unintended consequences?*
- o *Would inclusion of this measure in the Medicaid Adult Core Set be reasonable? Are there other programs that might benefit from inclusion of this measure?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:
**Seems OK
**Not in use.
**The measure needs to be applied to provider groups and treatment programs to begin to see if the metric does differentiate provider capability. This measure does not address efficiency of care. Assumption are made that treatment adherence is the proxy for quality and efficiency. Measure needs to be paired to a measurement of functional status change.
**It is not currently in use. It seems usable if the issues above are addressed.

Criterion 5: Related and Competing Measures

Related measures

- 0004: Initiation and Engagement of Alcohol and Other Drug Dependence Treatment (IET)
- 1664: SUB-3 Alcohol & Other Drug Use Disorder Treatment Provided or Offered at Discharge and SUB-3a Alcohol & Other Drug Use Disorder Treatment at Discharge

Harmonization

- Measure #0004 was discussed with the Behavioral Health Standing Committee in October, 2016. Since that time, the developer has continued its internal re-evaluation of the measure and may provide updates (if any are available at this time).

- Measure #1664 is a facility-level measure for the hospital setting. Differences in denominator definitions (i.e., including individuals with either alcohol or drug use disorder; different age ranges) will be discussed.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because empirical validity testing for the measure score has not been conducted and it has not been vetted by those being measured or others.

Pre-meeting public and member comments

- None received

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** NQF 3172
Measure Title: Continuity of Pharmacotherapy for Alcohol Use 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: 1/12/2017

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

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

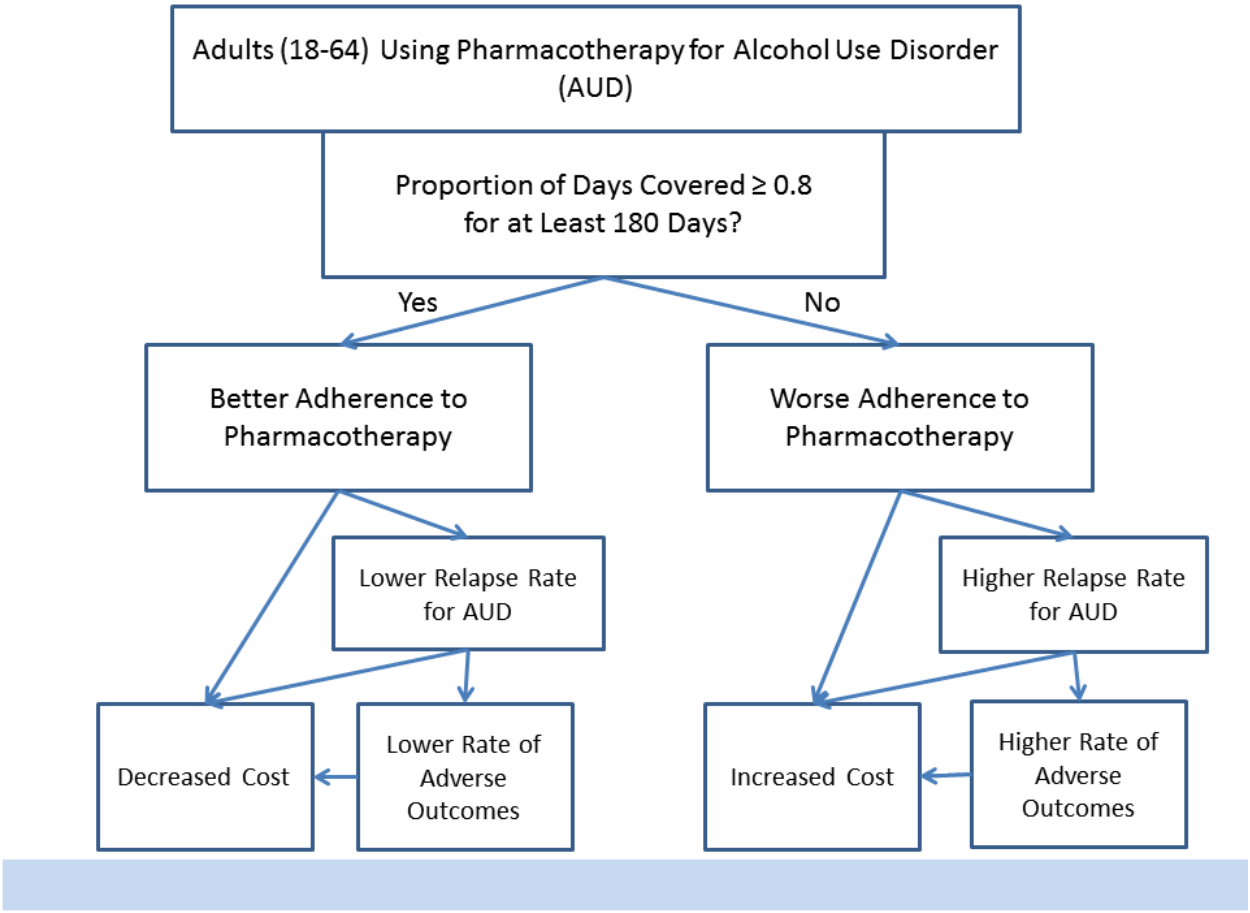
☒ Process: Continuity of Pharmacotherapy for Alcohol Use Disorder

☒ Appropriate use measure: Proportion of days covered by a medication for alcohol use disorder

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.12 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.



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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

The proposed measure is not an outcome measure.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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: See the section on "Justification of Measure Definition" in Section 1a.4 at the end of this file.

On the following pages, we present five tables (Exhibits 1-5) to summarize the evidence cited by the VA/DoD 2015 Guideline on Management of Substance Use Disorders to support the recommendations related to treatment of alcohol use disorder with pharmacotherapy. The tables contain paraphrased and verbatim excerpts from the systematic reviews, meta-analyses, and other studies, and are organized by type of medication:

- Exhibit 1: Acamprosate (systematic review; Jonas et al., 2014)
- Exhibit 2: Disulfiram (meta-analysis; Skinner et al., 2014)
- Exhibit 3: Naltrexone- oral or extended release (systematic review; Jonas et al., 2014)
- Exhibit 4: Topiramate (meta-analysis; Blodgett et al., 2014)
- Exhibit 5: Gabapentin (two randomized controlled trials; Anton et al., 2011 and Mason et al., 2014)

In Section 1a.4 (Other Source of Evidence) below, we include another source of evidence, under the heading, "Justification of Measure Definition". Under this heading, we include a summary of evidence that supports the measure definition. This material also appears in Section S.5. (Numerator Details) of the Measure Information Form (MIF).

Exhibit 1. Systematic Review Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Acamprosate for Alcohol Use Disorders (Jonas et al., 2014)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Systematic Review of Evidence Related to Treatment of Alcohol Use Disorder with Acamprosate:** Jonas DE, Amick HR, Feltner C, Bobashev G, Thomas K, Wines R, Kim MM, Shanahan E, Gass CE, Rowe CJ, et al. Pharmacotherapy for adults with alcohol use disorders in outpatient settings: a systematic review and metaanalysis. JAMA. 2014;311(18):1889–900. Cited in support of Recommendation 5 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 5 (Section D.a.i, page 33) from 2015 VA/DoD Guideline** For patients with moderate-severe alcohol use disorder, we recommend offering one of the following medications: • **Acamprosate** • Disulfiram • Naltrexone- oral or extended release • Topiramate

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Grades assigned to the evidence are as follows:
	• Acamprosate for adults with alcohol use disorder was significantly associated with improvements in return to any drinking (Strength of Evidence= Moderate)
	• Acamprosate for adults with alcohol use disorder was not significantly associated with improvement in return to heavy drinking (Strength of Evidence= Moderate)
	• No statistically significant difference was found for adults with alcohol use disorder between acamprosate and naltrexone for:
	  • Return to any drinking (Strength of Evidence= Moderate)
	  • Return to heavy drinking (Strength of Evidence= Moderate).
	Details are provided below in the section called, “Estimates of benefit and consistency across studies.”

Provide all other grades and definitions from the evidence grading system	Jonas et al., 2014: “We graded the strength of evidence as high, moderate, low, or insufficient based on established guidance [Owens et al., 2010]. The approach incorporates 4 key domains: risk of bias, consistency, directness, and precision. Two reviewers assessed each domain for each outcome and determined an overall grade. Differences were resolved by consensus.” Owens et al., 2010: Strength of Evidence grades and definitions: High=High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect. Moderate=Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate.

	Low=Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and is likely to change the estimate. Insufficient=Evidence either is unavailable or does not permit a conclusion. Citation: Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions: Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol. 2010;63(5):513-523.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 5 was "Strong For." "A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes." (page 11) "Using these elements, the grade of each recommendation is presented as part of a continuum: "Strong For (or "We recommend offering this option …") " (page 11)

Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation:
	• Balance of desirable and undesirable outcomes
	• Confidence in the quality of the evidence
	• Patient or provider values and preferences
	• Other implications, as appropriate, e.g.,:
	  • Resource use
	  • Equity
	  • Acceptability
	  • Feasibility
	  • Subgroup considerations
	Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation.
	
	They also determined the direction of each recommendation (For or Against). 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 …")

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	One systematic review and meta-analysis of 27 studies (n=7,519) assessed acamprosate's benefits and harms for adults with alcohol use disorder. The review included studies that enrolled adults with alcohol use disorders that evaluated FDA-approved medications for at least 12 weeks in an outpatient setting. Double blind randomized trials comparing acamprosate to placebo or another medication and prospective cohort studies that compared 2 medications were included. For adverse events, nonrandomized or open label trials, subgroup analyses from trials, prospective cohort studies, and case-control studies were included if they compared drugs of interest. The average age of patients in the studies included was usually in the 40's.

	Limitations: Only trials that were at least 12 weeks of treatment were included. The review did not assess how medication and psychosocial interventions compare with each other. The review also combined studies that included populations with a dual diagnosis and those that did not. The review also noted that publication bias and selective reporting are always potential limitations.
Estimates of benefit and consistency across studies	Acamprosate for adults with alcohol use disorder was associated with improvements in consumption outcomes generally. Return to any drinking: To prevent one person from returning to drinking for adults with alcohol use disorder, the NNT for acamprosate was 12 (95% CI, 8 to 26; 16 trials; n=4,847), RD=-0.09 (95% CI, -0.14 to -0.04). The RD for adults with alcohol use disorder in the 16 trials ranged from -0.27 (95% CI, -0.39 to -0.14) to 0.12 (95% CI, –0.00 to 0.25) (Strength of Evidence= Moderate). Subgroup analyses did show a reduction in effect size as the risk of bias of the study decreased: high/unclear RD=-0.13 (95% CI, -0.20 to -0.06; 3 trials; n=757), medium RD=-0.11 (95% CI, -0.16 to -0.06; 12 trials; n=3,438), low RD=-0.02 (95% CI, -0.09 to 0.05; 4 trials; n=1,409). Return to heavy drinking: Acamprosate was not associated with improvement in return to heavy drinking for adults with alcohol use disorder RD=-0.01 (95% CI, -0.04 to 0.03), ranging in the 7 trials from -0.13 (95% CI, -0.33 to 0.08) to 0.04 (95% CI, -0.09 to 0.16) (Strength of Evidence= Moderate). Meta-analyses of trials comparing acamprosate to naltrexone for adults with alcohol use disorder found no statistically significant difference between them for return to any drinking (RD, 0.02; 95% CI, −0.03 to 0.08 in 3 trials; n=800) (Strength of Evidence= Moderate) or heavy drinking (RD, 0.01; 95% CI, −0.05 to 0.06 in 4 trials; n=1,141) (Strength of Evidence= Moderate). * NNT is Number Needed to Treat, RR is Relative Risk, RD is Risk Difference
What harms were identified?	Patients treated with acamprosate had a higher risk of anxiety (NNH=7; 95% CI, 5 to 11; 2 trials; n=624), diarrhea (NNH=11; 95% CI, 6 to 34; 12 trials; n=2,978), vomiting (NNH=42; 95% CI, 24 to 143; 4 trials; n=1,817) than those on placebo. * NNH is Number Needed to Harm
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new studies were identified.

Exhibit 2. Meta-Analysis Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Disulfiram for Alcohol Use Disorders (Skinner et al., 2014)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Meta-Analysis of Evidence Related to Treatment of Alcohol Use Disorder with Disulfiram:** Skinner MD, Lahmek P, Pham H, Aubin HJ. Disulfiram efficacy in the treatment of alcohol dependence: A meta-analysis. PLoS One. 2014;9(2):e87366. Cited in support of Recommendation 5 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 5 (Section D.a.i, page 33) from 2015 VA/DoD Guideline** For patients with moderate-severe alcohol use disorder, we recommend offering one of the following medications: • Acamprosate • **Disulfiram** • Naltrexone- oral or extended release • Topiramate
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The authors did not assign a grade to the evidence supporting the recommendation for disulfiram. However, the article states, "The methodological quality of the studies was analyzed according to the Cochrane Collaboration's tool for assessing risk of bias. The 6-item tool assesses the quality of the randomization procedure (adequate sequence generation and allocation concealment), the blinding of treatments, the probability of other bias, the probability of selective reporting, and issues of incomplete data." The maximum possible score on the tool is meeting all of the six criteria. "Among the 22 studies included [in the meta-analysis], two met four of these six criteria, two met three criteria, eleven studies met two criteria, one met one criterion, and six studies met none of them (Table 2)." The average score for the 22 studies was 1.68 out of the six criteria.
Provide all other grades and definitions from the evidence grading system	See above row.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 5 was "Strong For." "A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes." (page 11) "Using these elements, the grade of each recommendation is presented as part of a continuum: "Strong For (or "We recommend offering this option …") " (page 11)
Provide all other grades and definitions from	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the

the recommendation grading system	strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation:
	• Balance of desirable and undesirable outcomes
	• Confidence in the quality of the evidence
	• Patient or provider values and preferences
	• Other implications, as appropriate, e.g.,:
	  • Resource use
	  • Equity
	  • Acceptability
	  • Feasibility
	  • Subgroup considerations
	Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation.
	
	They also determined the direction of each recommendation (For or Against). 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 …")

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	One meta-analysis of 22 RCTs that study disulfiram use with alcoholics in comparison to an alcoholic control group (n=2,414). The majority of those in the study were men (n=2058, 88.6%). Two studies included adolescent patients (average age 17). The studies ranged in years published from 1973 to 2010. Limitations: Most subjects in the analysis were men. In three studies, disulfiram was taken along with another treatment. Self-reporting was used for compliance in supervised studies. Trial methodologies were diverse because of the long period over which this treatment has been available. A large number of open trials were included as blinding disulfiram treatment is not always suitable.
Estimates of benefit and consistency across studies	"Overall, the 22 included studies showed a higher success rate of disulfiram compared to controls (Hedges'g = 0.58; 95% CI = 0.35–0.82).The effect size in the 22 studies ranged from -0.595 (95% CI, -1.279 to 0.089) to 1.965 (95% CI, 0.207 to 3.723). When comparing blind and open-label RCTs, only open-label trials showed a significant superiority over controls (g = 0.70; 95%CI = 0.46–0.93), because the drug's effectiveness depends directly on the patient's anticipations of being sick if alcohol is consumed, and therefore to observe effectiveness requires an open-label design. Disulfiram was also more effective than the control condition when compared to naltrexone (g = 0.77; 95% CI = 0.52–1.02), to acamprosate (g = 0.76; 95% CI = 0.04–1.48), and to the no disulfiram groups (g = 0.43; 95% CI = 0.17–0.69)."
What harms were identified?	Disulfiram was associated with increased risk of any adverse events compared with controls (RR=1.40; 95% CI, 1.01 to 1.94) in the 14 studies that reported them. The effect size in the 14 studies ranged from 0.156 (95% CI, 0.061 to 0.401) to 30.526 (95% CI, 1.821 to 511.722) (n=962) (e.g., hospitalization, death). * RR is Relative Risk

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new studies were identified.

Exhibit 3. Systematic Review Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Naltrexone for Alcohol Use Disorders (Jonas et al., 2014)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Systematic review of Evidence Related to Treatment of Alcohol Use Disorder with Naltrexone (oral or extended release):** Jonas DE, Amick HR, Feltner C, Bobashev G, Thomas K, Wines R, Kim MM, Shanahan E, Gass CE, Rowe CJ, et al. Pharmacotherapy for adults with alcohol use disorders in outpatient settings: a systematic review and metaanalysis. JAMA. 2014;311(18):1889–900. Cited in support of Recommendation 5 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 5 (Section D.a.i, page 33) from 2015 VA/DoD Guideline** For patients with moderate-severe alcohol use disorder, we recommend offering one of the following medications: • Acamprosate • Disulfiram • **Naltrexone- oral or extended release** • Topiramate

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Grades assigned to the evidence for naltrexone (oral or extended release) in adults with alcohol use disorder cited to support the recommendation are as follows:
	
	• Oral naltrexone (50 mg/d) for adults with alcohol use disorder was significantly associated with improvements in return to any drinking (Strength of Evidence= Moderate)
	• No significant association with return to any drinking was found for oral naltrexone (100 mg/d) (Strength of Evidence= Low) or injectable naltrexone (Strength of Evidence= Low) for adults with alcohol use disorder.
	• Oral naltrexone (50 mg/d) for adults with alcohol use disorder was significantly associated with improvements in return to heavy drinking (Strength of Evidence= Moderate)
	• No significant association with return to heavy drinking was found for oral naltrexone (100 mg/d) (Strength of Evidence= Low) or injectable naltrexone (Strength of Evidence= Low) for adults with alcohol use disorder.
	• Injectable naltrexone was significantly associated with reductions in heavy drinking days (Strength of Evidence= Low) for adults with alcohol use disorder.
	• No statistically significant difference was found between acamprosate and naltrexone for adults with alcohol use disorder in:
	  • Return to any drinking (Strength of Evidence= Moderate)
	  • Return to heavy drinking (Strength of Evidence= Moderate).

	Details are provided below in the section called, "Estimates of benefit and consistency across studies."
Provide all other grades and definitions from the evidence grading system	"We graded the strength of evidence as high, moderate, low, or insufficient based on established guidance [Owens et al., 2010]. The approach incorporates 4 key domains: risk of bias, consistency, directness, and precision. Two reviewers assessed each domain for each outcome and determined an overall grade. Differences were resolved by consensus." Strength of Evidence grades and definitions (Owens et al., 2010): High=High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect. Moderate=Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate. Low=Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and is likely to change the estimate. Insufficient=Evidence either is unavailable or does not permit a conclusion. Citation: Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions: Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol. 2010;63(5):513-523.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 5 was "Strong For." "A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes." (page 11) "Using these elements, the grade of each recommendation is presented as part of a continuum: "Strong For (or "We recommend offering this option …") " (page 11)

Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation:
	• Balance of desirable and undesirable outcomes
	• Confidence in the quality of the evidence
	• Patient or provider values and preferences
	• Other implications, as appropriate, e.g.,:
	  • Resource use
	  • Equity
	  • Acceptability
	  • Feasibility
	  • Subgroup considerations
	Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation.
	
	They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	One systematic review and meta-analysis of 53 studies (n=9,140) that assessed naltrexone's benefits and harms for adults with alcohol use disorder. The review included studies that enrolled adults with alcohol use disorders that evaluated FDA-approved medications for at least 12 weeks in an outpatient setting. Double blind randomized trials comparing naltrexone to placebo (N=44) or another medication (N=9) and prospective cohort studies that compared 2 medications were included. For adverse events, nonrandomized or open label trials, subgroup analyses from trials, prospective cohort studies, and case-control studies were included if they compared drugs of interest. The average age of patients in the studies included was usually in the 40's. Limitations: Only trials that were at least 12 weeks of treatment were included. The review did not assess how medication and psychosocial interventions compare with each other. The review also combined studies that included populations with a dual diagnosis and those that did not. The review also noted that publication bias and selective reporting are always potential limitations.
Estimates of benefit and consistency across studies	Oral naltrexone for adults with alcohol use disorder was associated with improvements in consumption outcomes generally. Return to drinking: To prevent 1 person from returning to any drinking, the NNT for oral naltrexone (50 mg/d) was 20 (95% CI, 11 to 500; 16 trials; n=2,347), RD=-0.05 (95% CI, -0.10 to -0.002), ranging in the 16 trials from -0.28 (95% CI, -0.44 to -0.11) to 0.10 (95% CI, -0.05 to 0.25) (Strength of Evidence= Moderate). No such association was found for oral naltrexone (100 mg/d) (Strength of Evidence=Low) or injectable naltrexone (Strength of Evidence= Low). Return to heavy drinking: To prevent 1 person from returning to heavy drinking, the NNT for oral naltrexone (50 mg/d) was 12 (95% CI, 8 to 26; 19 trials; n=2,875), RD=-0.09 (95% CI, -0.13 to -0.04), ranging in the 19 trials from -0.26 (95% CI, -0.44 to -0.09) to 0.08 (95% CI, -0.13 to 0.28) (Strength of Evidence= Moderate). No such association was found for oral naltrexone (100 mg/d) (Strength of Evidence= Low) or injectable naltrexone(Strength of Evidence= Low). Reduction in heavy drinking days: Injectable naltrexone was associated with reductions in heavy drinking days WMD=-4.6% (95% CI, -8.5% to -0.56%; 2 trials; n=926) (Strength of Evidence= Low). Meta-analyses of trials comparing naltrexone to acamprosate found no statistically significant difference between them for return to any drinking (RD, 0.02; 95%CI, −0.03 to 0.08 in 3 trials; n=800) (Strength of Evidence= Moderate) or heavy drinking (RD, 0.01; 95%CI, −0.05 to 0.06 in 4 trials; n=1,141) (Strength of Evidence= Moderate). * NNT is Number Needed to Treat, RR is Relative Risk, WMD is Weighted Mean Difference
What harms were identified?	Patients treated with naltrexone had a higher risk of withdrawal from trials due to adverse events (NNH=48; 95% CI, 30 to 112; 17 trials; n=2,743), and had a higher risk of dizziness (NNH=16; 95% CI, 12 to 28; 13 trials; n=2,675), nausea (NNH=9; 95% CI, 7 to 14; 24 trials; n=4,655) and vomiting (NNH=24; 95% CI, 17 to 44; 9 trials; n=2,438) than those on placebo. * NNH is Number Needed to Harm
Identify any new studies conducted since the SR. Do the new studies change the	No new studies were identified.

conclusions from the SR?	

Exhibit 4. Meta-Analysis Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Topirimate for Alcohol Use Disorders (Blodgett et al., 2014)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Meta-Analysis of Evidence Related to Treatment of Alcohol Use Disorder with Topirimate:** Blodgett JC, Del Re AC, Maisel NC, Finney JW. A meta-analysis of topiramate's effects for individuals with alcohol use disorders. Alcohol Clin Exp Res. Jun 2014;38(6):1481-1488. Cited in support of Recommendation 5 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 5 (Section D.a.i, page 33) from 2015 VA/DoD Guideline** For patients with moderate-severe alcohol use disorder, we recommend offering one of the following medications: • Acamprosate • Disulfiram • Naltrexone- oral or extended release • **Topiramate**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The meta-analysis did not assign a grade to the evidence, but states, "It provides an up-to-date estimate of the efficacy of topiramate for treating AUDs, focusing on high quality (randomized, double-blind, placebo-controlled) trials."
Provide all other grades and definitions from the evidence grading system	The meta-analysis did not assign a grade to the evidence.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 5 was "Strong For." "A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes." (page 11) "Using these elements, the grade of each recommendation is presented as part of a continuum: "Strong For (or "We recommend offering this option …") " (page 11)
Provide all other grades and definitions from	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the

the recommendation grading system	strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation:
	• Balance of desirable and undesirable outcomes
	• Confidence in the quality of the evidence
	• Patient or provider values and preferences
	• Other implications, as appropriate, e.g.,:
	  • Resource use
	  • Equity
	  • Acceptability
	  • Feasibility
	  • Subgroup considerations
	Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation.
	
	They also determined the direction of each recommendation (For or Against). 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 …")

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: One meta-analysis of 7 RCTs (n=1,125) that compared topiramate to placebo for treatment for people with alcohol use disorder and one RCT (n=30) of flexible-dose topiramate on 30 veterans with PTSD and alcohol use disorder. Quality of studies: All included studies were randomized, double-blind, placebo-controlled. The main limitation is the small number of studies included. The effect sizes reported should be read in context of the small number of studies included. Studies outside of this review have studied low-dose topiramate (up to 75 mg), while this study only looks at dosages of topiramate of 250 mg.
Estimates of benefit and consistency across studies	In the seven studies, topiramate's effects on drinking outcomes were favorable, and small to moderate in magnitude. Abstinence: Hedge's g=0.468 (95% CI, 0.250 to 0.687) (p<0.01), ranging from 0.056 (95% CI, -0.473 to 0.585) to 0.774 (95% CI, 0.597 to 0.950). Heavy Drinking: Hedges' g=0.406 (95% CI, 0.215 to 0.600) (p<0.01), ranging from 0.140 (95% CI, -0.114 to 0.395) to 0.623 (95% CI, 0.345 to 0.901). Cravings: Hedge's g=0.312 (95% CI, -0.042 to 0.666) (P=0.07), ranging from -0.172 (95% CI, -0.583 to 0.240) to 0.736 (95% CI, 0.226 to 1.245).
What harms were identified?	In the four studies that looked at drop out due to adverse event, 42/342 participants (12.3%) dropped out on topiramate versus 16/348 (4.6%) on placebo. Overall dropout rates were similar between topiramate and placebo (37.2% versus 38.1%) and were not significant in an omnibus test across five studies with reported information.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Additional study: Batki SL, Pennington DL, Lasher B, et al. Topiramate treatment of alcohol use disorder in veterans with posttraumatic stress disorder: A randomized controlled pilot trial. Alcohol Clin Exp Res. Aug 2014;38(8):2169-2177. Although this study has a small sample size and is limited to patients with a comorbid diagnosis of PTSD, the results support the conclusion that topirimate decreases the percentage of drinking days. One double blind, placebo-controlled, randomized pilot trial (n=30) of flexible-dose topiramate in 30 veterans with PTSD and alcohol use disorder over 12 weeks. Limitations: The trial had a small sample size which did not allow for the examination of other factors that may have influenced the outcomes, such as moderating effects of concomitant treatment, genetics, degree of motivation at study entry, or presence of pretreatment abstinence. Self-reporting measures to assess outcomes were also used. <u>Estimates of benefit and consistency across studies</u> The group taking topiramate showed a significant decrease in the percentage of drinking days from baseline through week 12 (RR=0.89, 95% CI 0.82-0.98; p=0.019), and when compared to a placebo group,exhibited a larger decrease in drinking days that was nearly significant (RR=0.430; 95% CI, 0.18 to 1.05; p=0.063). Topiramate treatment revealed a trend for reduced standard drinks in a week (55%; p=0.099) and reduced drinks per drinking day (61%; p=0.057) during the course of treatment weeks 1-12 compared to the placebo group. Topiramate treatment showed a significant reduction in alcohol craving using the Obsessive Compulsive Disorder Scale (OCDS) from baseline to week 12 (p=0.002). There was also a larger reduction in craving in the topiramate group when compared to the placebo group (p=0.025). <u>What harms were identified?</u> There were no significant differences between the topiramate and the placebo groups in the rate of adverse events. The most common reported adverse events were sleepiness, loss of appetite, change of sense of taste, itching, diarrhea, and abnormal vision. * RR is Relative Risk

Exhibit 5. Two Randomized Controlled Trials Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Gabapentin for Alcohol Use Disorders (Anton et al., 2011 and Mason et al., 2014)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**Two Randomized Controlled Trials Related to Treatment of Alcohol Use Disorder with Gabapentin:** Anton RF, Myrick H, Wright TM, et al. Gabapentin combined with naltrexone for the treatment of alcohol dependence. Am J Psychiatry. 2011 Jul;168(7):709-17. PMID: 21454917. Mason BJ, Quello S, Goodell V, Shadan F, Kyle M, Begovic A. Gabapentin treatment for alcohol dependence: A randomized clinical trial. JAMA Intern Med. Jan 2014;174(1):70-77. Cited in support of Recommendation 6 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 6 (Section D.a.i, page 33) from 2015 VA/DoD Guideline** For patients with moderate-severe alcohol use disorder for whom first-line pharmacotherapy is contraindicated or ineffective, we suggest offering **gabapentin**.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The VA/DoD Guideline did not assign a grade to the evidence they cited to support Recommendation 6.
Provide all other grades and definitions from the evidence grading system	The VA/DoD Guideline did not assign a grade to the evidence they cited to support Recommendation 6.
Grade assigned to the **recommendation**	The grade assigned to Recommendation 6 was "Weak For." "If the Work Group is less confident of the balance between desirable and undesirable outcomes, they give a weak recommendation." (page 11)

with definition of the grade	"Using these elements, the grade of each recommendation is presented as part of a continuum: "Weak For (or "We suggest offering this option …")." (page 11)

Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation:
	• Balance of desirable and undesirable outcomes
	• Confidence in the quality of the evidence
	• Patient or provider values and preferences
	• Other implications, as appropriate, e.g.,:
	  • Resource use
	  • Equity
	  • Acceptability
	  • Feasibility
	  • Subgroup considerations
	Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation.
	
	They also determined the direction of each recommendation (For or Against). 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 …")

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: Two RCTs have been published looking at the efficacy of treating Alcohol Use Disorder (AUD) with gabapentin. The first study (Anton et al., 2011) uses a double dummy placebo controlled medication design to randomize 150 patients with alcohol dependence into three groups: naltrexone/gabapentin, naltrexone-only, and double placebo to look at the efficacy of gabapentin in conjunction with naltrexone versus naltrexone-only and placebo. The mean age was in the mid-40's, primarily male and Caucasian. The article flagged several limitations of the study, including it is a single site study with a limited number of individuals who do not have other significant psychiatric conditions, are not on psychiatric medications, are medically stable, and are motivated about abstaining. The independent effect of gabapentin could not be estimated (Anton et al., 2011). The second study (Mason et al., 2014) is a double-blind, placebo-controlled, randomized dose-ranging trial of 150 men and women older than 18 years old with alcohol use disorder to observe whether gabapentin improves outcomes in a dose-dependent way. The dosages were 0 mg (placebo), 900 mg/d, 1800mg/d. All analysis was intent to treat. The participants were in their early to mid-40's.The study flagged several limitations; including a significant drop out rate after assessment and that it was a single-site study (Mason et al., 2014).

Estimates of benefit and consistency across studies	**Anton 2011:** Time to first heavy drink day: During the first 6 weeks when subjects received gabapentin, the naltrexone/gabapentin group had a longer time to relapse than the naltrexone-alone group (p=0.04). The naltrexone-alone group and the placebo group had no statistically significant difference in this measure. There was no difference between groups for the remainder of the trial (10 additional weeks). Percent heavy drinking days: During the first 6 weeks, a difference was not observed between the naltrexone/gabapentin and placebo groups. The naltrexone/gabapentin did better than the naltrexone-alone group (p=0.0002). Drinks per drinking day: During the first 6 weeks, the naltrexone/gabapentin group did significantly better than the naltrexone-only group (p=0.02) and the placebo group (p=0.01). Cravings: Using the Obsessive Compulsive Drinking Scale (OCDS), there was no significant difference between groups in either phase of the study. In the subscale of the OCDS called the resistance control index (RCI), naltrexone/gabapentin showed a significantly lower score (more control of drinking urges) than the naltrexone-only group (p=0.04) and somewhat lower than the placebo group, although not significant. Sleep quality: During the first 6 weeks, the naltrexone/gabapentin group experienced significantly better sleep quality than the placebo group (p=0.02) and the naltrexone-only group (p=0.03). **Mason 2013:** Rate of complete abstinence: Over the 12-week treatment course, gabapentin had a significant linear dose effect on increasing rates of complete abstinence (p=0.04). Placebo had a sustained abstinence rate of 4.1% (95% CI, 1.1-13.7), 900-mg gabapentin had a sustained abstinence rate of 11.1% (95% CI, 5.2-22.2), and 1800-mg gabapentin had a sustained abstinence rate of 17.0% (95% CI, 8.9-30.1). The 1800-mg group had an OR=4.8 (95% CI, 0.9-35.0), compared to the placebo group. Rate of no heavy drinking: Over the 12-week treatment course, gabapentin had a significant linear dose effect on increasing rates of complete abstinence (p=0.02). Placebo had a rate of no heavy drinking of 22.5% (95% CI, 13.6-37.2), 900-mg gabapentin had a rate of no heavy drinking of 29.6% (95% CI, 19.1-42.8), and 1800-mg gabapentin had a rate of no heavy drinking of 44.7% (95% CI, 31.4-58.8). The 1800-mg group had an OR=2.8 (95% CI, 1.1-7.5) compared to the placebo group. Average number of days of heavy drinking per week: Those taking increasing doses of gabapentin showed a significant linear decrease in average number of days of heavy drinking per week (p<0.001). In the 900-mg group, the reduction was 1.8 days (95% CI, -2.2 to -1.3) (p<0.001) and in the 1800-mg group, the reduction was 2.0 days (95% CI, -2.5 to -1.5) (p<0.001) Number of drinks consumed per week: Those taking increasing doses of gabapentin showed a significant linear decrease in average number of days of heavy drinking per week (p<0.001). In the 900-mg group, the reduction of 2.2 drinks was not significant (95% CI, -5.3 to 1.0) and in the 1800-mg group, the reduction was 6.7 drinks (95% CI, -9.8 to -3.5) (p<0.001). Cravings: Gabapentin showed significant linear dose effects on cravings using the Alcohol Craving Questionnaire (p=0.03). Comparing 1800-mg gabapentin to placebo, there was a significant reduction (6.8 [95% CI, -12.1 to -1.5)]; p=0.01).

	Mood: Gabapentin showed significant linear dose effects on mood using the Beck Depression Inventory (p=0.001). Comparing 1800-mg gabapentin to placebo, there was a significant reduction (1.1 [95% CI, -2.0 to -0.3)]; p=0.01). Quality of Sleep: Gabapentin showed significant linear dose effects on sleep using the Pittsburgh Sleep Quality Index total score (p<0.001). Comparing 1800-mg gabapentin to placebo, there was a significant reduction (1.5 [95% CI, -2.1 to –0.8)]; p<0.001). OR=Odds Ratio
What harms were identified?	**Anton 2011:** The naltrexone/gabapentin and the naltrexone-only groups showed more dizziness than placebo (p=0.006). The naltrexone/gabapentin group showed more daytime sleepiness (p=0.02) than both of the other groups, and more blurred vision (p=0.02) and more premature ejaculation (p=0.02) than the placebo group. **Mason 2013:** No differences were found among groups in rates of adverse events. Groups were found to have a similar number and severity of reported adverse events (i.e., fatigue, insomnia, headache).
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new studies were identified.

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.

In this section, we provide evidence that justifies the measure definition. We also present this evidence in the Numerator Details section of the MIF (see Section S.5).

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

Justification of Measure Definition: Most experts believe that substance dependence, including AUD, should be considered and treated as a chronic illness (McLellan, 2000). This is particularly important for patients who seek and receive treatment for alcohol dependence (McLellan, 2002), as they have self-identified as being at higher risk for a chronic relapsing course (Dawson, 1996; Institute of Medicine, 1998; McKay, 2005; Miller & Hester, 1986). Similar to other chronic diseases, patients may experience relapses, treatment is optimized when patients remain in continuing care, and outcomes are better in patients who receive treatment for longer compared to shorter durations (Moos et al., 1999; NIDA, 1999; Ouimette et al., 1998).

We define treatment continuity as (1) receiving at least 180 days of treatment and (2) medication available for at least 80% of the days of treatment.

Our definition of minimum duration is based on the fact that the FDA-registration trials for AUD drugs studied the effect of treatment over three to six months (US FDAa, undated; US FDAb, undated), and we have no evidence for effectiveness of shorter durations. In addition, the greatest risk of relapse is in the first 6-12 months after alcohol abstinence is initiated (US FDAa, undated; US FDAb, undated; US DHHS, 2015; Medical Services Commission, 2011). We did not specify a maximum duration of treatment, as no upper limit for duration of treatment has been empirically established (US DHHS, 2015).

Our definition of adherence follows the established convention of a 0.8 threshold for proportion of days covered (PDC) (Seabury et al., 2015). This threshold was used, for example, to assess naltrexone adherence in a study of AUD treatment (Kranzler et al., 2008). Sufficient treatment adherence is essential, as medication non-adherence is associated with relapse to heavy and/or frequent drinking and higher health care utilization (Gueorguieva et al., 2013; Kranzler et al., 2008; Stout et al., 2014).

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

We conducted a targeted literature search supplemented by a manual search of the references cited in relevant articles.

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

Dawson DA. Correlates of past-year status among treated and untreated persons with former alcohol dependence: United States, 1992. Alcoholism, Clinical and Experimental Research. 1996;20(4):771-779.

Gueorguieva R, Wu R, Krystal JH, Donovan D, O'Malley SS. Temporal patterns of adherence to medications and behavioral treatment and their relationship to patient characteristics and treatment response. Addictive Behaviors. 2013;38:2119-21.

Institute of Medicine. Bridging the Gap Between Practice and Research: Forging Partnerships with Community-Based Drug and Alcohol Treatment. Washington, DC: The National Academies Press; 1998.

Kranzler HR, Stephenson JJ, Montejano L, Wang S, Gastfried DR. Persistence with oral naltrexone for alcohol treatment: implications for health-care utilization. Addiction. 2008;103:1801-1808.

McKay JR. Is there a case for extended interventions for alcohol and drug use disorders? Addiction. 2005;100(11):1594-1610.

McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. JAMA. 2000;284(13):1689-95.

McLellan AT. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. Addiction. 2002;97(3):249-252.

Medical Services Commission, British Columbia: Problem Drinking (2011). Accessed November 23 at: http://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/bc-guidelines/problem-drinking

Miller WR, Hester RK. The Effectiveness of Alcoholism Treatment. In: Miller WR, Heather N, eds. Treating Addictive Behaviors: Processes of Change. Boston, MA: Springer US; 1986:121-174.

Moos RH, Finney JW, Ouimette PC, Suchinsky RT. A comparative evaluation of substance abuse treatment: I. Treatment orientation, amount of care, and 1-year outcomes. Alcohol Clin Exp Res. 1999;23(3):529-36.

National Institute on Drug Abuse (NIDA). Principles of Drug Addiction Treatment: A Research-Based Guide. NIH Publication No. 99–4180. Rockville, MD: NIDA, 1999, reprinted 2000.

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522.

Seabury SA, Lakdawalla DN, Dougherty JS, Sullivan J, Goldman DP. Medication Adherence and Measures of Health Plan Quality. Am J Manag Care. 2015;21(6):e379-e389

Stout RL, Braciszewski JM, Subbaraman MS, Kranzler HR, O'Malley SS, Falk D. What happens when people discontinue taking medications? Lessons from COMBINE. Addiction. 2014;109:2044-2052.

U.S. Department of Health and Human Services (DHHS) Assistant Secretary for Planning and Evaluation Office of Disability, Aging and Long-Term Care Policy. Review of Medication-Assisted Treatment Guidelines and Measures for Opioid and Alcohol Use. Washington, DC, 2015. Accessed November 9, 2016 at: https://aspe.hhs.gov/sites/default/files/pdf/205171/MATguidelines.pdf

U.S. Food and Drug Administration (FDA) (a). REVIA Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/018932s017lbl.pdf

U.S. Food and Drug Administration (FDA) (b). VIVITROL Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/021897lbl.pdf

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_3172_AUD_Evidence_Form_1-12-17_To_NQF.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

In spite of its high prevalence and its substantial burden on patients, their families and society, alcohol use disorder (AUD) remains a severely undertreated condition. According to the 2014 National Survey on Drug Use and Health (NSDUH), 16.3 million Americans ages 18 years and older suffered from AUD (SAMHSA, 2014a), representing almost 7 percent of the adult population (SAMHSA, 2014b). But only 15.2 percent of patients, who reported that they needed alcohol treatment, actually received it (SAMHSA, 2014c).

Medication-assisted treatment (i.e., pharmacotherapy combined with counseling) is an evidence-based and effective treatment option for patients with AUD. It is supported by several national guidelines and a Consensus Panel of the National Institute on Alcohol Abuse and Alcoholism and the Substance Abuse and Mental Health Services Administration (SAMHSA, 2015), but substantially underused. A recent study that included 16,947 AUD patients treated in a network of community health centers, for example, found that only 3.2 percent of those patients received pharmacotherapy (Riekmann et al., 2016). Minority patients and those without health insurance were significantly less likely to receive pharmacotherapy.

For patients receiving pharmacotherapy, continuity of treatment is critical. Most experts believe that substance dependence, including AUD, should be considered and treated as a chronic illness (McLellan, 2000). This is particularly important for those patients who seek and receive treatment (McLellan, 2002), as they have self-identified to be at risk for a chronic relapsing course (Dawson, 1996; Institute of Medicine, 1998; McKay, 2005; Miller & Hester, 1986).

Overall, longer duration of AUD treatment is associated with better outcomes (Lemke & Moos, 2003; Moos et al., 1995; Oimette et al., 1998) and treatment adherence is essential, as medication non-adherence is associated with relapse to heavy and/or frequent drinking and higher health care utilization (Gueorguieva et al., 2013; Kranzler et al., 2008; Stout et al., 2014). Yet evidence suggests that persistence of pharmacotherapy is poor. Baser et al. looked at medication adherence for six months after initiation of treatment in a large database study including 15,502 patients treated with an FDA-approved AUD drug. They found that only six to eleven percent of patients on oral drugs and only 21 percent of patients with injectable naltrexone were sufficiently adherent (Baser et al., 2011).

Therefore, the proposed measure focuses on continuity of pharmacotherapy, defined as treatment duration of at least 180 days and sufficient adherence for the duration of treatment. The definition of adherence follows the established convention of having access to medication for at least 80 percent of treatment days.

Several important benefits related to quality improvement are envisioned with the implementation of this measure. First, the measure will help health plans and providers to identify individuals with AUD, who are non-adherent to or discontinue pharmacotherapy. As a result, this measure will encourage health plans and providers to develop communication and education tools and processes to improve treatment continuity in their patients with AUD. Improved treatment continuity is expected to result in lower rates of relapse, and less substance use-related morbidity and mortality. Adoption of this performance measure has the potential to improve quality of care for individuals with AUD and, therefore, advance quality of care by engaging patients as partners in their care, and promoting effective communication and coordination of care, priority areas identified in the National Quality Strategy.

CITATIONS

Baser O, Chalk M, Rawson R, Gastfriend DR. Alcohol dependence treatments: comprehensive healthcare costs, utilization outcomes, and pharmacotherapy persistence. Am J Manag Care [2011, 17 Suppl 8:S222-34].

Dawson DA. Correlates of past-year status among treated and untreated persons with former alcohol dependence: United States, 1992. Alcoholism, Clinical and Experimental Research. 1996;20(4):771-779.

Gueorguieva R, Wu R, Krystal JH, Donovan D, O'Malley SS. Temporal patterns of adherence to medications and behavioral treatment and their relationship to patient characteristics and treatment response. Addictive Behaviors. 2013;38:2119-21.

Institute of Medicine. Bridging the Gap Between Practice and Research: Forging Partnerships with Community-Based Drug and Alcohol Treatment. Washington, DC: The National Academies Press; 1998.

Kranzler HR, Stephenson JJ, Montejano L, Wang S, Gastfried DR. Persistence with oral naltrexone for alcohol treatment: implications for health-care utilization. Addiction. 2008;103:1801-1808.

Lemke S, Moos RH. (2003). Outcomes at 1 and 5 years for older patients with alcohol use disorders. J Subst Abuse Treat. 24(1):43-50.

McKay JR. Is there a case for extended interventions for alcohol and drug use disorders? Addiction. 2005;100(11):1594-1610.

McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. JAMA. 2000;284(13):1689-95.

McLellan AT. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. Addiction. 2002;97(3):249-252.

Miller WR, Hester RK. The Effectiveness of Alcoholism Treatment. In: Miller WR, Heather N, eds. Treating Addictive Behaviors: Processes of Change. Boston, MA: Springer US; 1986:121-174.

Moos RH, Pettit B, Gruber V. (1995). Longer episodes of community residential care reduce substance abuse patients´ readmission rates. J Stud Alcohol. 56(4):433-43.

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522.

Riekmann T, Muench J, McBurnie M, et al. Medication-assisted treatment for substance use disorders within a national community health center research network. Subst Abuse. 2016;37(4):625-634.

Stout RL, Braciszewski JM, Subbaraman MS, Kranzler HR, O'Malley SS, Falk D. What happens when people discontinue taking medications? Lessons from COMBINE. Addiction. 2014;109:2044-2052.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014a). 2014 National Survey on Drug Use and Health (NSDUH). Table 5.8A—Substance dependence or abuse in the past year among persons aged 18 or older, by demographic characteristics: Numbers in thousands, 2013 and 2014. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.htm#tab5-8a

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014b). 2014 National Survey on Drug Use and Health (NSDUH). Table 5.8B—Substance dependence or abuse in the past year among persons aged 18 or older, by demographic characteristics: Percentages, 2013 and 2014. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.htm#tab5-8b

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014c). 2014 National Survey on Drug Use and Health (NSDUH). Table 5.24A – Locations Received Alcohol Treatment in the Past Year among Persons Who Received Alcohol Treatment in the Past Year, by Age Group: Numbers in Thousands, 2013 and 2014. Available at: http://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs2014/NSDUH-DetTabs2014.htm#tab5-24a

Substance Abuse and Mental Health Services Administration and National Institute on Alcohol Abuse and Alcoholism (SAMHSA). (2015). Medication for the Treatment of Alcohol Use Disorder: A Brief Guide. HHS Publication No. (SMA) 15-4907. Rockville, MD: Substance Abuse and Mental Health Services Administration.

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 (4b) under Usability and Use.*

The measure scores are calculated based on two-year rolling periods for 2010-2015. The overall measure score results are shown in Table 1 and the results by state and health plan are shown in Tables 2 and 3, respectively. The number of patients in the denominator of the measure ranged from 11,737 in 2010-2011 to 26,803 in 2014-2015 (Table 1).

Over the period from 2010-2015, measure scores increased from 0.162 to 0.214. Over the 2010-2015 time period, the number of states with at least 20 eligible patients in the denominator increased from 44 states in 2010-2011 to 47 states in 2014-2015 (Table 2). Over the 2010-2015 time period, the number of health plans with at least 20 eligible patients in the denominator increased from 58 plans in 2010-2011 to 203 plans in 2014-2015 (Table 3).

Table 1. Denominator, Numerator, and Measure Score for Two-Year Rolling Periods, 2010-2015

Time Period	Denominator	Numerator	Score
2010-2011	11,737	1,907	0.162
2011-2012	17,381	3,302	0.190
2012-2013	19,273	3,743	0.194
2013-2014	22,330	4,305	0.193
2014-2015	26,803	5,732	0.214

Table 2. Summary Statistics for Measure Scores by State, Two-Year Rolling Periods, 2010-2015

Time Period	Number of States	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
2010-2011	44	0.161	0.162	0.067	0.241	0.035	0.045	0.121	0.138	0.162	0.183	0.200
2011-2012	45	0.194	0.187	0.087	0.300	0.041	0.052	0.149	0.172	0.187	0.224	0.237
2012-2013	45	0.196	0.194	0.062	0.278	0.044	0.048	0.151	0.175	0.194	0.222	0.250
2013-2014	46	0.195	0.197	0.116	0.305	0.036	0.037	0.152	0.170	0.197	0.206	0.237
2014-2015	47	0.218	0.211	0.125	0.364	0.036	0.033	0.187	0.202	0.211	0.235	0.256

Table 3. Summary Statistics for Measure Scores by Health Plan, Two-Year Rolling Periods, 2010-2015

Time Period	Number of Health Plans	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
2010-2011	58	0.149	0.146	0	0.350	0.078	0.100	0.045	0.091	0.146	0.190	0.250
2011-2012	93	0.194	0.200	0	0.409	0.080	0.112	0.091	0.138	0.200	0.250	0.300
2012-2013	138	0.192	0.180	0	0.440	0.074	0.097	0.103	0.143	0.180	0.240	0.304
2013-2014	179	0.195	0.192	0	0.422	0.075	0.083	0.100	0.150	0.192	0.233	0.302
2014-2015	203	0.219	0.221	0.048	0.421	0.073	0.110	0.125	0.161	0.221	0.271	0.313

STD=standard deviation; IQR=interquartile range; PNN=NNth percentile

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.**

Data from the testing of the measure as specified are provided in 1b.2 above.

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 (4b) under Usability and Use.*

The measure was stratified for disparities by age and gender. The results/scores are presented for these categories in Table 4. For ease of exposition, we only present data for 2013-2014, the period for which we have the most data.

Table 4. Scores by Age and Gender for Entire Sample, 2013-2014

Category	Denominator	Numerator	Measure Score
Age			
18-64 years	22,330	4,305	0.193
18 –34	6,238	936	0.150
35 –44	4,775	856	0.179
45 –54	6,656	1,368	0.206
55 – 64	4,661	1,145	0.246
Gender			
Both	22,330	4,305	0.193
Female	10,392	2,172	0.209
Male	11,938	2,133	0.179

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**

Data on disparities from the testing of the measure as specified are provided in 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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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.)*

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_3172_AUD_Code_Lists_1-12-17_To_NQF.xlsx

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

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

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Individuals in the denominator who have at least 180 days of treatment and a PDC of at least 0.8 for AUD 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 measure numerator is calculated based on administrative claims data for rolling two-year periods from 2010 to 2015: 2010-2011, 2011-2012, 2012-2013, 2013-2014, and 2014-2015. The numerator of the measure is defined as individuals with a PDC of 0.8 or greater for an AUD medication over a minimum of 180 days. To meet the 180-day requirement and be eligible for the PDC, the date on the first claim for an AUD medication must fall at least 180 days before the end of the final calendar year of the measurement period. The PDC is calculated as follows:

PDC NUMERATOR

The PDC numerator is the sum of the days covered by the days' supply of all claims for all oral AUD medications plus 30 days for each extended-release injection of AUD medication. For claims with a days' supply that extends beyond the end of the measurement period, count only the days for which the medication was available (or last effective date for injections) during the measurement period. If two or more prescription claims occur on the same day or overlap, the surplus based on the days' supplies accumulates over all prescriptions. If there are two or more claims for an injectable AUD medication for the same drug (generic name) on the same date of service, keep the claim with the largest days' supply. However, if another claim is submitted after an injectable AUD medication claim, the surplus from the days' supply for the initial injectable AUD medication claim is not retained.

PDC DENOMINATOR

The PDC denominator is the number of days from the first claim date for an AUD medication through the last day of the days' supply of the last claim (or last effective date for injections) during the measurement period or the end of the measurement period, whichever comes first. The period of time covered by the PDC starts on the day of the first claim (index date) and must last for a minimum of 180 days even if there are no additional claims for AUD medication. To meet the 180-day requirement and be eligible for the PDC, the date on the first claim for an AUD medication must fall at least 180 days before the end of the measurement period.

AUD medications were identified using National Drug Codes (NDCs) for the following:

- Acamprosate
- Disulfiram
- Naltrexone (oral)
- Topiramate
- Gabapentin

Or a HCPCS code for the following injectable medication:

- Naltrexone (extended-release injectable)

We have included topiramate and gabapentin because they are recommended in the 2015 VA/DoD Guideline. Although they are not FDA-approved for treatment of alcohol use disorder, supporting evidence exists to justify their inclusion (see Evidence Form).

The NDCs for the oral medications and the HCPCS code for the injectable medication are contained in the sheets called "NDCs" and "HCPCS Codes", respectively, in the Excel file called "NQF 3172 AUD Code Lists" which is attached to this form under Item S.2b.

Justification of Measure Definition: Most experts believe that substance dependence, including AUD, should be considered and treated as a chronic illness (McLellan, 2000). This is particularly important for patients who seek and receive treatment for alcohol dependence (McLellan, 2002), as they have self-identified to be at higher risk for a chronic relapsing course (Dawson, 1996; Institute of Medicine, 1998; McKay, 2005; Miller & Hester, 1986). Similar to other chronic diseases, patients may experience relapses, treatment is optimized when patients remain in continuing care, and outcomes are better in patients who receive treatment for longer compared to shorter durations (Moos et al., 1999; NIDA, 1999; Ouimette et al., 1998).

We define treatment continuity as (1) receiving at least 180 days of treatment and (2) medication available for at least 80% of the days of treatment.

Our definition of minimum duration is based on the fact that the FDA-registration trials for AUD drugs studied the effect of treatment over three to six months (US FDAa, undated; US FDAb, undated), and we have no evidence for effectiveness of shorter durations. In addition, the greatest risk of relapse is in the first 6-12 months after alcohol abstinence is initiated (US FDAa, undated; US FDAb, undated; US DHHS, 2015; Medical Services Commission, 2011). We did not specify a maximum duration of treatment, as no upper limit for duration of treatment has been empirically established (US DHHS, 2015).

Our definition of adherence follows the established convention of a 0.8 threshold for proportion of days covered (PDC) (Seabury et al., 2015). This threshold was used, for example, to assess naltrexone adherence in a study of AUD treatment (Kranzler et al.,

2008). Sufficient treatment adherence is essential, as medication non-adherence is associated with relapse to heavy and/or frequent drinking and higher health care utilization (Gueorguieva et al., 2013; Kranzler et al., 2008; Stout et al., 2014).

CITATIONS

Dawson DA. Correlates of past-year status among treated and untreated persons with former alcohol dependence: United States, 1992. Alcoholism, Clinical and Experimental Research. 1996;20(4):771-779.

Gueorguieva R, Wu R, Krystal JH, Donovan D, O'Malley SS. Temporal patterns of adherence to medications and behavioral treatment and their relationship to patient characteristics and treatment response. Addictive Behaviors. 2013;38:2119-21.

Institute of Medicine. Bridging the Gap Between Practice and Research: Forging Partnerships with Community-Based Drug and Alcohol Treatment. Washington, DC: The National Academies Press; 1998.

Kranzler HR, Stephenson JJ, Montejano L, Wang S, Gastfried DR. Persistence with oral naltrexone for alcohol treatment: implications for health-care utilization. Addiction. 2008;103:1801-1808.

McKay JR. Is there a case for extended interventions for alcohol and drug use disorders? Addiction. 2005;100(11):1594-1610.

McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. JAMA. 2000;284(13):1689-95.

McLellan AT. Have we evaluated addiction treatment correctly? Implications from a chronic care perspective. Addiction. 2002;97(3):249-252.

Medical Services Commission, British Columbia: Problem Drinking (2011). Accessed November 23 at: http://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/bc-guidelines/problem-drinking

Miller WR, Hester RK. The Effectiveness of Alcoholism Treatment. In: Miller WR, Heather N, eds. Treating Addictive Behaviors: Processes of Change. Boston, MA: Springer US; 1986:121-174.

Moos RH, Finney JW, Ouimette PC, Suchinsky RT. A comparative evaluation of substance abuse treatment: I. Treatment orientation, amount of care, and 1-year outcomes. Alcohol Clin Exp Res. 1999;23(3):529-36.

National Institute on Drug Abuse (NIDA). Principles of Drug Addiction Treatment: A Research-Based Guide. NIH Publication No. 99–4180. Rockville, MD: NIDA, 1999, reprinted 2000.

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522.

Seabury SA, Lakdawalla DN, Dougherty JS, Sullivan J, Goldman DP. Medication Adherence and Measures of Health Plan Quality. Am J Manag Care. 2015;21(6):e379-e389

Stout RL, Braciszewski JM, Subbaraman MS, Kranzler HR, O'Malley SS, Falk D. What happens when people discontinue taking medications? Lessons from COMBINE. Addiction. 2014;109:2044-2052.

U.S. Department of Health and Human Services (DHHS) Assistant Secretary for Planning and Evaluation Office of Disability, Aging and Long-Term Care Policy. Review of Medication-Assisted Treatment Guidelines and Measures for Opioid and Alcohol Use. Washington, DC, 2015. Accessed November 9, 2016 at: https://aspe.hhs.gov/sites/default/files/pdf/205171/MATguidelines.pdf

U.S. Food and Drug Administration (FDA) (a). REVIA Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/018932s017lbl.pdf

U.S. Food and Drug Administration (FDA) (b). VIVITROL Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/021897lbl.pdf

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Individuals 18-64 years of age who had a diagnosis of AUD and at least one claim for an AUD 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).
The measure denominator is calculated for rolling two-year periods from 2010 to 2015: 2010-2011, 2011-2012, 2012-2013, 2013-2014, and 2014-2015. The denominator includes individuals 18-64 years of age during their treatment period who had a diagnosis code of AUD during an inpatient, intensive outpatient, partial hospitalization, outpatient, detoxification or emergency department encounter at any time during the measurement period. To meet the 180-day requirement and be eligible for the measure, the date on the first claim for an AUD medication must fall at least 180 days before the end of the measurement period.

The diagnosis codes used to identify individuals with AUD included:
- ICD-9: 291.xx, 303.xx, 305.0x
- ICD-10: F10.xxx

These codes with descriptions are contained in the sheets called "ICD-9 Diagnosis Codes" and "ICD-10 Diagnosis Codes" in the Excel file called "NQF 3172 AUD Code Lists" which is attached to this form under Item S.2b.

AUD medications were identified using National Drug Codes (NDCs) for the following:
- Acamprosate
- Disulfiram
- Naltrexone (oral)
- Topiramate
- Gabapentin

Or the HCPCS code for the following injectable medication:
- Naltrexone (extended-release injectable)

We have included topiramate and gabapentin because they are recommended in the 2015 VA/DoD Guideline. Although they are not FDA-approved for treatment of alcohol use disorder, supporting evidence exists to justify their inclusion (see Evidence Form). The NDCs for the oral medications and the HCPCS code for the injectable are contained in the sheets called "NDCs" and "HCPCS Codes", respectively, in the Excel file called "NQF 3172 AUD Code Lists" which is attached to the form under Item S.2b.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
There are no denominator exclusions.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no denominator exclusions.

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.)*
Measure results may be stratified by:
- Age – Divided into four categories: 18-34, 35-44, 45-54, 55-64 years
- Gender: Male, Female

- State
- Health plan

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
The measure score is calculated for rolling two-year periods from 2010 to 2015. The steps described below are repeated for five two-year periods: 2010-2011, 2011-2012, 2012-2013, 2013-2014, and 2014-2015. We present detailed results in the MIF for 2013-2014, as we have the most data for this time period, but we include measure scores for each of the two-year periods within 2010-2015.

DENOMINATOR: Individuals 18-64 years of age who had a diagnosis of AUD and at least one claim for an AUD medication

CREATE DENOMINATOR:

1. For each two-year period, identify individuals who are 18-64 years of age for the duration of the first year during which they appear in the period.

2. Of individuals identified in Step 1, keep those who had at least one encounter with any diagnosis (primary or secondary) of AUD in an outpatient setting, acute inpatient setting, or emergency department setting at any time during the two-year measurement period. The AUD diagnosis codes with descriptions are contained in the sheets called “ICD-9 Diagnosis Codes” and “ICD-10 Diagnosis Codes” in the Excel file called “NQF 3172 AUD Code Lists”, which is attached to this form under Item S.2b.

3. Of individuals identified in Step 2, keep those who have at least one claim with a National Drug Code (NDC) for any of the following oral AUD medications during the two-year period:
- Acamprosate
- Disulfiram
- Naltrexone (oral)
- Topiramate
- Gabapentin

Or the HCPCS code for the following injectable AUD medication:
- Naltrexone (extended-release injectable)

Claims for oral medications with negative, missing, or zero days’ supply were not included. The NDCs for the oral medications and the HCPCS code for the injectable are contained in the sheets called “NDCs” and “HCPCS Codes”, respectively, in the Excel file called “NQF 3172 AUD Code Lists,” which is attached to this form under Item S.2b.

4. Of individuals identified in Step 3, keep individuals who were continuously enrolled in a commercial health plan captured by our data for at least 6 months after the month with the first AUD medication claim in the measurement period, with no gap in enrollment. Individuals who are not enrolled for 6 months, including those who die before 6 months of enrollment, are not eligible and are not included in the analysis. This is the denominator.

NUMERATOR: Individuals in the denominator who have at least 180 days of treatment and a PDC of at least 0.8 for AUD medications

CREATE NUMERATOR:

For the individuals in the denominator, calculate the PDC for each individual using the following method:

1. Determine the number of days for the PDC denominator. The start date is the service date (fill date) of the first prescription or injection claim for an AUD medication in the two-year measurement period. The end date is defined as the earliest of:

- The date on which the individual exhausts their days' supply, including any pre-existing surplus, following their final claim (assuming daily use).
- The individual's death date.
- December 31st of the second year in the two-year period.

2. For each individual: Of the days included in the PDC denominator, count those for which the individual was covered by at least one AUD medication based on the prescription drug or injection claim service date and days' supply, using the methods in 2a-2d below. This is the PDC numerator.

2a. Sort AUD medication claims by individual's ID and service date. Scan the claims in order, calculating a rolling surplus which accumulates any remaining supply from other prior or same-day fills.

2b. Naltrexone injections contribute 30 days' supply unless another claim is found sooner, in which case the Naltrexone injection covers only the days up to the next claim. Claims for Naltrexone injections are not added to the surplus supply and only one such claim per day is counted.

2c. For each individual, calculate the number of days within the PDC denominator that are covered by an AUD medication, including the surplus supply, as described in the previous step, as available medication.

2d. 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.

3. Calculate the PDC for each individual. Divide the number of days covered by an AUD medication (PDC numerator from Step 2) by the number of days in the PDC denominator (from Step 1).

4. Of the individuals in Step 3, count the number of individuals who have a period of 180 days or greater from the start date of the first claim for AUD medication to the end date of the last claim for AUD medication within the two-year period and have a calculated PDC of at least 0.8 for AUD medication during the period. This is the numerator.

CALCULATE MEASURE SCORE:

1. Calculate the measure score by dividing the numerator by the denominator.

2. Calculate the measure score for each state. The state code on the claim record is used to identify individuals in each state. The measure score is then reported for each state that has at least 20 individuals in the denominator.

3. Calculate the measure score for each health plan. Health plan membership is approximated based on a combination of industry type and Metropolitan Statistical Area (MSA). A health plan identifier is assigned based on each unique combination of industry and MSA. The health plan identifier is used to group individuals into health plans. The measure score is then reported for each health plan that has at least 20 individuals in the denominator.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF a PRO-PM, identify whether (and how) proxy responses are allowed.

Not applicable; 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.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable; this measure does not use a sample or survey.

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

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
For measure calculation, the following files from the Truven MarketScan® Commercial Database were used:

- Enrollment data
- Drug claims
- Medical claims

We used data from these files (including data from Standard Quarterly Updates) for calendar years 2010-2015. This database has long been a commonly used data source to study patterns of commercially insured patients. The database contains fully adjudicated, patient-level claims. All records in these files were used as input to identify individuals that met the measure's eligibility criteria. We present detailed results in the MIF for 2013-2014, as we have the most data for this time period, but we include measure scores for each of the two-year periods within 2010-2015. The final analytic file for 2013-2014 contained a total of 22,330 episodes.

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, Population : Regional and State

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health : Outpatient, Clinician Office/Clinic
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
NQF_3172_AUD_Testing_Form_1-12-17_To_NQF.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. (Do not remove prior testing information – include date of new information in red.)

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. (Do not remove prior testing information – include date of new information in red.)

2.3 For maintenance of endorsement

Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

Measure Number (*if previously endorsed*)**:** 3172
Measure Title: Continuity of Pharmacotherapy for Alcohol Use Disorder
Date of Submission: 1/12/2017
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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

For measure calculation, the following files from the Truven MarketScan® Commercial Database were used:

- Enrollment data
- Drug claims
- Medical claims

We used data from these files (including data from Standard Quarterly Updates) for calendar years 2010-2015 for two-year rolling measurement periods. This database has long been a commonly used data source to study patterns of commercially insured patients. The database contains fully adjudicated, patient-level claims. All records in these files were used as input to identify individuals that met the measure's eligibility criteria. We present detailed results in the MIF and this testing form for 2013-2014, as we have the most data for this time period, but we include measure scores for each of the two-year periods within 2010-2015. We restricted the sample to members with continuous enrollment of at least 180 days. The final analytic file for 2013-2014 contained a total of 22,330 episodes.

1.3. What are the dates of the data used in testing? January 1, 2010 – December 31, 2015

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.26*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: 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*)

Characteristics of the 2013-2014 denominator are summarized in Table 1 for the two levels of analysis, state and health plan. The sample for the state analysis included 46 states with 20 or more members eligible for the denominator.

As the data do not contain an actual health plan identifier, we developed a method based on the fact that the claims data are sourced from self-insured employers. We approximated health plan membership based on a combination of variables for industry type and Metropolitan Statistical Area (MSA) and assigned identifiers based on each unique

combination of industry and MSA. The sample for the health plan analysis included 179 health plans with 20 or more members eligible for the denominator.

Table 1. Demographic Characteristics by States and Health Plans, 2013-2014

	States (n=46)	Health Plans (n=179)
Mean number per unit	471	33
Median number per unit	278	38
Minimum number per unit	24	20
Maximum number per unit	2187	237
Standard Deviation	486.3	29.5
P10	54	21
P25	112	24
P50	278	33
P75	698	50
P90	1037	73

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*)

Demographic characteristics of the individuals in the 2013-2014 dataset are shown in Table 2. . For both the state and health plan analyses, the episodes are fairly equally distributed across the four age groups, and slightly more than half were male.

Table 2. Number of Individuals Included in Testing of AUD Measure for States and Health Plans, by Demographic Characteristics, 2013-2014

Characteristic	States (n=46)	Health Plans (n=179)
Total Population	21,674	7,631
Age	Percent	Percent
18-34	27.9	29.8
35-44	21.4	21.7
45-54	29.8	29.3
55-64	20.9	19.3
Gender		
Female	46.5	46.9
Male	53.5	53.2

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.

There were no differences in the data used for different aspects of testing (e.g., measure scores, reliability).

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

The patient-level sociodemographic (SDS) variables that were available and analyzed in the data were age and gender.

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*)

The method of reliability testing used and the rationale are described below.

Method of Reliability Testing and Rationale

In order to assess measure precision in the context of the observed variability across measurement units (states and health plans), 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 (2009): "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."

This approach has two basic assumptions:

1. Each measured entity has a true pass rate, p, which varies between units of measurement following an unknown distribution between 0 and 1; and,
2. The measured entity's score is a sample proportion, calculated from a binomial random sample conditional on the measured entity's true pass rate.

As defined by Adams (2009), signal is defined as the variance in true pass rate between units of measurement, noise is defined as the estimation variance for each measured entity's score, and reliability scores are a ratio of signal to the sum of signal and noise. We used the robust Prasad-Rao estimator for estimating the signal, and the standard binomial distribution inference for estimating the noise.

Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual unit variance); whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across units). 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 units of observation.

Calculations were based on the mean denominator size for states (n=471) and health plans (n=33). 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."

Only health plans and states with more than one patient in the denominator were included in the calculation since units with only one observation cannot show any within-unit variation.

CITATIONS

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*)

Reliability scores were calculated at the state level and health plan level for the 2013-2014 data.

The reliability score and standard deviation at the state level are 0.772 and 0.068, respectively, for the 2013-2014 data. This reliability score is greater than 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between states.

The reliability score and standard deviation at the health plan level are 0.846 and 0.053, respectively, for the 2013-2014 data. This reliability score is greater than 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between health plans.

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 results indicated that the measure scores were reliable at the state and health plan level, with both sets of measure scores having a reliability score of greater than 0.7, which is considered an acceptable cutpoint for adequate reliability (Scholle et al., 2008).

2b2. VALIDITY TESTING

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

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

☐ **Performance measure score**

☐ **Empirical validity testing**

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

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

We identified ten clinical experts in the treatment of AUD to rate the measure's face validity and usability using a web-

based questionnaire (developed using SurveyMonkey®). The names and organizations of the clinical experts are listed in Table 3.

The clinical experts were asked to review the measure specifications and the evidence supporting the measure from the NQF forms, which we provided to them. After reviewing the background material, they were instructed to rate two statements about the measure by indicating their level of agreement on a 5-point scale (1=Strongly Disagree; 2=Disagree; 3=Neither Agree nor Disagree; 4=Agree; 5=Strongly Agree).

The first statement was related to the face validity of the measure: "Performance scores resulting from the measure as defined can be used to distinguish good from poor quality." The second statement was related to the usability of the measure: "The measure results are easily understood by the users of the data (e.g., clinicians, administrators)."

Table 3. Clinical Experts Who Rated Measure on Face Validity and Usability

Name	Affiliations and Employment
Adam Bisaga, MD New York, NY	Professor of Psychiatry Columbia University College of Physicians & Surgeons
Mady Chalk, PhD, MSW Philadelphia, PA	Managing Director, The Chalk Group Senior Policy Advisor, Treatment Research Institute, Philadelphia, PA
Bowen Chung, MD, MSHS Santa Monica, CA	Attending Physician Department of Psychiatry, Harbor-UCLA Medical Center Psychiatrist County of Los Angeles Department of Mental Health Associate Professor-in-Residence, Department of Psychiatry and Bio-behavioral Sciences David Geffen School of Medicine at UCLA Adjunct Scientist RAND Corporation
Louisa Degenhardt, PhD Sydney, Australia	Professor of Epidemiology NHMRC Principal Research Fellow Fellow of the Academy of Social Sciences of Australia
Keith Heinzerling, MD, MPH Los Angeles, CA	Associate Professor in Residence UCLA Department of Family Medicine
Brian Hurley, MD, MBA, DFASAM Los Angeles, CA	Addiction Psychiatrist Treasurer, American Society of Addiction Medicine Los Angeles County Department of Mental Health - Robert Wood Johnson Foundation Clinical Scholar at the David Geffen School of Medicine of the University of California, Los Angeles
Richard Saitz, MD, MPH, FACP, DFASAM Boston, MA	Chair, Department of Community Health Sciences (CHS) Professor of Community Health Sciences & Medicine Boston University School of Public Health
Jeffrey Samet, MD, MA, MPH Boston, MA	Professor of Medicine & Community Health Sciences Boston University Schools of Medicine & Public Health John Noble MD Professor in General Internal Medicine & Professor of Public Health Chief, General Internal Medicine, Boston Medical Center
Andrew Saxon, MD Seattle, WA	Professor and Director, Addiction Psychiatry Residency Program Department of Psychiatry & Behavioral Sciences University of Washington Director, Center of Excellence in Substance Abuse Treatment and Education (CESATE) VA Puget Sound Health Care System

Name	Affiliations and Employment
Constance Weisner, DrPH, MSW Oakland, CA	Professor, Department of Psychiatry University of California San Francisco Research Scientist, Division of Research Kaiser Permanente Northern California

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

Ten experts completed the evaluation of the measure's face validity and usability. The results of the rating of face validity on a scale of 1 to 5 are presented in Table 4.

Table 4. Results of the Face Validity Evaluation

Rating	Number with Rating (%)
5 (Strongly Agree)	2 (20%)
4 (Agree)	4 (40%)
3 (Neither Agree nor Disagree)	4 (40%)
2 (Disagree)	0 (0%)
1 (Strongly Disagree)	0 (0%)

Of the experts who rated the measure for face validity, 60 percent (6/10) strongly agreed or agreed with this statement: "Performance scores resulting from the measure as defined can be used to distinguish good from poor quality". The mean rating for face validity was 3.8, and the median rating 4.

The results of the rating of usability on a scale of 1 to 5 are presented in Table 5.

Table 5. Results of the Usability Evaluation

Rating	Number with Rating (%)
5 (Strongly Agree)	4 (40%)
4 (Agree)	3 (30%)
3 (Neither Agree nor Disagree)	1 (10%)
2 (Disagree)	2 (20%)
1 (Strongly Disagree)	0 (0%)

Of the experts who rated the measure for usability, 70 percent (7/10) strongly agreed or agreed with this statement: "The measure results are easily understood by the users of the data (e.g., clinicians, administrators)." The mean rating for usability was 3.9, and the median rating 4.

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

In summary, 60 percent of the alcohol use disorder experts who participated in the rating strongly agreed or agreed that the measure has face validity, and 70 percent strongly agreed or agreed that the measure exhibits usability. This indicates sufficient support for the validity and usability of the measure.

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

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
Not applicable

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
Not applicable

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
Not applicable

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

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

2b4.1.1 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

2b4.2. 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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
Not applicable

2b4.4a. What were the statistical results of the analyses used to select risk factors?
Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable

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

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

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable

2b4.9. Results of Risk Stratification Analysis:
Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not applicable

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not applicable

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

To identify statistically significant differences in performance, we conducted a comparison of means and percentiles at the state and health plan level. Confidence intervals (95% CI) were calculated around point estimates for each state and health plan and then compared to the overall mean of states and health plans, respectively. If the confidence intervals did not overlap with the overall mean, the difference was considered statistically significant.

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

We analyzed the 2013-2014 measure scores by state and health plan. The results for measure scores by state and health plan are presented in Table 6, along with a discussion of the meaningful differences at each level.

Table 6. Measure Score Performance at the State and Health Plan Level, 2013-2014

Level	n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
State	46	0.195	0.197	0.116	0.305	0.036	0.037	0.152	0.170	0.197	0.206	0.237
Health Plan	179	0.195	0.192	0	0.422	0.075	0.083	0.100	0.150	0.192	0.233	0.302

Meaningful Differences at the State Level – 2013-2014

In 2013-2014, 3 of the 46 states (6.5 percent) had scores statistically significantly lower than the state-level mean, and 3 states (6.5 percent) had scores significantly higher than the state-level mean. For states with at least 20 episodes, state-level measure scores ranged from a minimum of 0.116 to a maximum of 0.305, indicating suboptimal performance across all 46 states.

Meaningful Differences at the Health Plan Level – 2013-2014

In 2013-2014, 12 of 179 health plans (6.7 percent) were statistically significantly lower than the health plan-level mean, and 5 (2.8 percent) of health plans were statistically significantly higher than the health plan-level mean. For those health plans with at least 20 episodes, the scores for the low- (10^{th} percentile) - and high (90^{th} percentile) performing plans were 0.100 and 0.302, respectively, indicating suboptimal performance across all plans and wide variation between low- and high-performing plans.

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

The results indicate that overall measure performance is suboptimal with variation in performance across states and health plans. Statistically significant differences were identified at the state and health plan level when compared to the overall mean.

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Not applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

One possible threat to validity is missing days' supply, which is a required data element to calculate the measure. An empirical assessment of this was conducted by analyzing the number (%) of individuals in a measure denominator in 2010-2015 with one or more claims that had missing days' supply.

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

Of 62,309 individuals in at least one of the 2010-2015 AUD cohorts, just 457 (0.7%) had one or more drug claims with a negative, zero, or missing value for days' supply. This small number of cases indicates that missing data do not pose a threat to the validity of the measure.

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

Threats to validity from missing data, to the extent we were able to address with testing, were not identified. The findings from the exploratory analysis suggest that very little impact on measure rates would be expected from 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.**
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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
The measure is not in operational use. Testing demonstrated that the measure was feasible to specify and calculate using administrative claims data. The claims data needed to implement the measure are available, accessible, and timely. Issues affecting feasibility regarding missing data were not identified. The cost of data collection is negligible, since the administrative data (collected primarily for billing purposes) are used as the data source for this measure. No other feasibility/implementation issues were identified.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees, licensing, or other requirements to use any aspect of the measure as specified.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.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 being submitted for initial endorsement.

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

Not applicable; the measure is being submitted for initial endorsement.

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

Because the measure is being submitted to NQF for initial endorsement, we have not yet submitted it for use in a specific federal, state or local program. However, this measure would be appropriate for use in a Centers for Medicare & Medicaid Services (CMS) reporting program for Medicaid patients, such as the 2016 Core Set of Behavioral Health Measures for Medicaid. This list of 13 behavioral health measures was identified by CMS for voluntary reporting by state Medicaid and Children´s Health Insurance Program (CHIP) agencies. We will explore the possibility of submitting this measure through the Measures under Consideration (MUC) process for the one of the CMS reporting programs. This would entail submitting information about the measure through

JIRA, which is the CMS software system for collecting information on candidate measures for the list of "Measures under Consideration" for the annual pre-rulemaking process.

Improvement

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

4b. **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.
Not applicable; information about progress on improvement is not required because this measure is being submitted for initial endorsement.

4c. Unintended Consequences

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

4c.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
The measure has not been implemented in any reporting programs, and unexpected positive or negative findings were not identified during testing.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
The measure has not been implemented in any reporting programs, and therefore, unexpected benefits from implementation have not been observed.

4d1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Not applicable; this measure is being submitted for initial endorsement.

4d1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
Not applicable; this measure is being submitted for initial endorsement.

4d2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
Not applicable; this measure is being submitted for initial endorsement.

4d2.2. **Summarize the feedback obtained from those being measured.**
Not applicable; this measure is being submitted for initial endorsement.

4d2.3. **Summarize the feedback obtained from other users**
Not applicable; this measure is being submitted for initial endorsement.

4d.3. **Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**

Not applicable; this measure is being submitted for initial endorsement.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

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

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

1664 : SUB-3 Alcohol & Other Drug Use Disorder Treatment Provided or Offered at Discharge and SUB-3a Alcohol & Other Drug Use Disorder Treatment at Discharge

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?

No

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

The target population of the proposed measure is related to the two measures listed above (NQF 0004 and NQF 1664). Differences among the three measures, along with the rationale and impact are discussed below in the text box for Item 5b.1. The text box for this item (5a.2) would not accommodate the length of our response.

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 that address both the same measure focus and the same target population.

RESPONSE TO ITEM 5A.2

The information below is the response to Item 5a.2 (previous item), describing the differences, rationale, and impact on interpretability and data collection burden for the two NQF-endorsed RELATED measures which were identified. (The text box under Item 5a.2 would not accept this volume of formatted text.)

The target population of the proposed measure is related to the two measures listed above (NQF 0004 and NQF 1664). The proposed measure focuses on continuity of pharmacotherapy for patients with AUD. NQF 0004 focuses on treatment initiation and engagement of patients with a new episode of AUD or other substance use disorders, including opioid use disorder (OUD). NQF 1664 focuses on AUD and other drug use disorders among hospital discharges. Differences among the three measures, along with the rationale and impact are discussed below.

Diagnoses included in denominator definition

- Proposed measure: diagnosis of AUD
- NQF 0004: diagnosis of alcohol or other drug dependence
- NQF 1664: diagnosis of AUD or another substance use disorder
- Rationale and impact of focusing on only AUD: There are different medications for treatment of AUD and OUD, and there are no FDA-approved medications for treatment of other substance use disorders. In addition, the conceptual issues related to continuity of pharmacotherapy differ between AUD and OUD, so developing separate measures for the two disorders is required. The impact of this is a more narrowly focused measure that provides information specific to individuals with AUD.

Age range

- Proposed measure: Patients 18-64 years of age
- NQF 0004: Patients 13 years of age and older
- NQF 1664: Patients 18 years of age and older
- Rationale and impact of limiting to individuals 18-64 years of age: Medications for treatment of AUD have not been approved by the FDA for adolescent patients 13-17 years of age; therefore, the proposed measure is restricted to adults 18-64 years of age.

Data Source

- Proposed measure: Electronic claims data
- NQF 0004: Administrative claims, electronic clinical data
- NQF 1664: Electronic clinical data, paper medical records
- Rationale and impact of using electronic claims data: Electronic claims data are timely, accessible, and relatively inexpensive to use for analyses of a large number of patients. Using a single source of data expedites the calculation of the measure, and will provide feedback to providers sooner.

Inpatient vs. outpatient

- Proposed measure: Inpatient and outpatient
- NQF 0004: Inpatient and outpatient
- NQF 1664: Inpatient discharges
- Rationale and impact of using inpatient and outpatient records to identify patients: A large majority of patients with AUD are not admitted to a hospital, so using inpatient and outpatient data leads to more complete identification of the population eligible for treatment.

Process of care included in numerator definition

- Proposed measure: Adherence to pharmacotherapy for AUD
- NQF 0004: Inpatient admission, outpatient visit, intensive outpatient encounter, or partial hospitalization for adults with a new episode of AUD, OUD, or other substance use disorders
- NQF 1664: Medication for treatment of alcohol or drug use disorder OR a referral for addictions treatment
- Rationale and impact of the process of care included in the numerator definition: Successful pharmacotherapy of AUD requires high adherence over at least a 180-day period. Therefore, providing feedback to providers about adherence to AUD treatment has the potential to improve the adherence rates by increasing provider awareness, and motivating health plans and insurers to develop educational material and programs about pharmacotherapy for AUD for both providers and patients.

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):** RAND Corporation
Co.2 **Point of Contact:** Soeren, Mattke, mattke@rand.org, 617-338-2059-8622
Co.3 **Measure Developer if different from Measure Steward:** RAND Corporation
Co.4 **Point of Contact:** Soeren, Mattke, mattke@rand.org, 617-338-2059-8622

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 group of ten behavioral health experts was used to rate the face validity and usability of the measure. Their names and affiliations are provided in the Testing Form.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Some proprietary codes are contained in the measure specifications for convenience of the user. Use of these codes may require permission from the code owner or agreement to a license.

Ad.7 **Disclaimers:** This performance measure does not establish a standard of medical care and has not been tested for all potential applications.

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 #: 3175
Corresponding Measures:
Measure Title: Continuity of Pharmacotherapy for Opioid Use Disorder
Measure Steward: RAND Corporation
Brief Description of Measure: Percentage of adults 18-64 years of age with pharmacotherapy for opioid use disorder (OUD) who have at least 180 days of continuous treatment
Developer Rationale: The rapidly rising number of deaths and near-deaths from opioid overdoses over the past several years has brought the issue of treating opioid use disorder (OUD) to the forefront of the policy agenda. The Surgeon General , who mailed a call to action to 2.3 million doctors, nurses, dentists, and other clinicians asking them to help address this escalating epidemic (Murthy, 2016), and the governors of many states have prioritized improving access to prevention and treatment of OUD.

The high prevalence of OUD has increased the sense of urgency: According to the 2014 National Survey on Drug Use and Health (NSDUH), 1.7 million adults 18 years and older were classified as having a pain reliever use disorder and 886,000 adults had used heroin in the past year (SAMHSA, 2015). In 2014, there were 489,532 episodes for OUD treatment, including outpatient treatment, detoxification, and residential treatment, in the Treatment Episode Data Set (TEDS) (SAMHSA, 2016).

Medication-assisted treatment (i.e., pharmacotherapy combined with counseling) is an evidence-based and effective treatment option for patients with OUD. Individuals with OUD receiving pharmacotherapy have significantly lower rates of mortality while they continue to receive medication, compared to individuals who were not receiving medication (Brugal et al., 2005; Cornish et al., 2010; Cousins et al., 2016; Davoli et al, 2007; Degenhardt et al., 2009; Gibson & Degenhardt, 2007; Pierce et al., 2016). But OUD medications remain markedly underutilized (Volkow et al., 2014). Of the almost half million aforementioned episodes, medication-assisted treatment was planned for only 25.4 percent (20.7 percent in outpatient treatment, 3.6 percent in detoxification, and 1.1 percent in residential treatment) (SAMHSA, 2016).

Additionally, ensuring treatment continuity is critical to the success of medication-assisted treatment. Longer treatment duration for individuals with OUD is associated with better outcomes and the best outcomes have been observed in patients in long-term methadone maintenance programs ("Effective medical treatment of opiate addiction", 1998; Moos et al., 1999; NIDA 1999; Ouimette et al., 1998; Peles et al., 2013).

In addition, there is strong evidence for increased mortality when individuals with OUD transition off pharmacotherapy, both in the short term (within the first 0-4 weeks after discontinuing pharmacotherapy) and over the long term (Cornish et al., 2010; Cousins et al., 2016; Davoli et al, 2007; Degenhardt et al., 2009; Gibson & Degenhardt, 2007; Pierce et al., 2016).

Therefore, the proposed measure focuses on continuity of pharmacotherapy, defined as treatment duration of at least 180 days and absence of treatment gaps of greater than 7 days. Several important benefits related to quality improvement are envisioned with the implementation of this measure. First, the measure will help health plans and providers to identify individuals with OUD who are non-adherent to or discontinue pharmacotherapy. As a result, this measure will encourage health plans and providers to develop communication and education tools and processes to improve treatment continuity in their patients with OUD. Improved treatment continuity is expected to result in lower rates of relapse, and less substance use-related morbidity and mortality. Adoption of this performance measure has the potential to improve quality of care for individuals with OUD and, therefore, advance quality of care by engaging patients as partners in their care, and promoting effective communication and coordination of care, priority areas identified in the National Quality Strategy.

CITATIONS

Brugal MT, Domingo-Salvany A, Puig R, et al. Evaluating the impact of methadone maintenance programmes on mortality due to overdose and aids in a cohort of heroin users in Spain. Addiction. 2005;100(7):981-989.

Cornish R, Macleod J, Strang J, Vickerman P, Hickman M. Risk of death during and after opiate substitution treatment in primary care: prospective observational study in UK General Practice Research Database. BMJ. 2010;341:c5475.

Cousins G, Boland F, Courtney B, Barry J, Lyons S, Fahey T. Risk of mortality on and off methadone substitution treatment in primary care: a national cohort study. Addiction. 2016;111(1):73-82.

Davoli M, Bargagli AM, Perucci CA, et al. Risk of fatal overdose during and after specialist drug treatment: the VEdeTTE study, a national multisite prospective cohort study. Addiction. 2007;102:1954-9.

Degenhardt L, Randall D, Hall W, Law M, Butler T, Burns L. Mortality among clients of a state-wide opioid pharmacotherapy program over 20 years: risk factors and lives saved. Drug and alcohol dependence. 2009;105:9-15.

Effective medical treatment of opiate addiction. National Consensus Development Panel on Effective Medical Treatment of Opiate Addiction. JAMA.1998;280:1936-1943.

Gibson AE, Degenhardt LJ. Mortality related to pharmacotherapies for opioid dependence: a comparative analysis of coronial records. Drug Alcohol Rev. 2007; 26(4), 405-410.

Moos RH, Finney JW, Ouimette PC, Suchinsky RT. A comparative evaluation of substance abuse treatment: I. Treatment orientation, amount of care, and 1-year outcomes. Alcohol Clin Exp Res. 1999;23(3):529-36.

Murthy VH. Ending the Opioid Epidemic - A Call to Action. N Engl J Med. 2016 Dec 22;375(25):2413-2415. doi: 10.1056/NEJMp1612578. Epub 2016 Nov 9.

National Institute on Drug Abuse (NIDA). Principles of Drug Addiction Treatment: A Research-Based Guide. NIH Publication No. 99–4180. Rockville, MD: NIDA, 1999, reprinted 2000.

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522.

Peles E, Schreiber S, Adelson M. Opiate-dependent patients on a waiting list for methadone maintenance treatment are at high risk for mortality until treatment entry. J Addict Med. 2013;7(3):177-82.

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, Jones A, et al. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. Addiction. 2016;111:298-308.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (2015). Behavioral health trends in the United States: results from the 2014 National Survey on Drug Use and Health. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2015: 7-12 (http://www .samhsa .gov/ data/sites/ default/ files/ NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf).

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (2016). Treatment Episode Data Set (TEDS): 2004-2014. National Admissions to Substance Abuse Treatment Services. BHSIS Series S-84, HHS Publication No. (SMA) 16-4986. Rockville, MD: Substance Abuse and Mental Health Services Administration. Available at http://www.samhsa.gov/data/sites/default/files/2014_Treatment_Episode_Data_Set_National_Admissions_9_19_16.pdf

Volkow ND, Frieden TR, Hyde PS, Cha SS. Medication-assisted therapies--tackling the opioid-overdose epidemic. N Engl J Med. 2014 May 29;370(22):2063-6

Numerator Statement: Individuals in the denominator who have at least 180 days of continuous pharmacotherapy with a medication prescribed for OUD without a gap of more than seven days
Denominator Statement: Individuals 18-64 years of age who had a diagnosis of OUD and at least one claim for an OUD medication
Denominator Exclusions: There are no denominator exclusions.

Measure Type: Process
Data Source: Claims (Other), Pharmacy
Level of Analysis: Health Plan, Population : Regional and State

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.
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 or **Summary of prior review in [year]**

- The developer provides a diagram of the relationship of this process of care (pharmacotherapy for opioid use disorder [OUD] for at least 180 days and absence of treatment gaps of greater than 7 days) to lower OUD relapse rates, which in turn leads to fewer adverse outcomes and decreased costs.
- Evidence provided by the developers to support the measure include recommendations from the VA/DoD 2015 Guideline on Management of Substance Use Disorders.

- **Recommendation 8:** For patients with opioid use disorder, we recommend offering one of the following medications considering patient preferences (***recommendation grade: "strong for…"***)
  - Systematic Review: Buprenorphine/naloxone (**evidence grade high-quality**)
    - Meta-analysis also provided.
  - Systematic Review: Methadone in an Opioid Treatment Program (**evidence grade high-quality**)
    - Meta-analysis also provided.
- **Recommendation 11:** For patients with opioid use disorder for whom opioid agonist treatment is contraindicated, unacceptable, unavailable, or discontinued and who have established abstinence for a sufficient period of time (see narrative), we recommend offering (***recommendation grade: "strong for…"***)
  - Extended-release injectable naltrexone (**not graded** - relied on one double-blind, placebo-controlled, randomized, 24-week trial of 250 patients)
- **Recommendation 12:** There is insufficient evidence to recommend for or against oral naltrexone for treatment of opioid use disorder. (Studies were of low quality and strength.)

- The developer provided evidence to support their use of the 180-day minimum treatment period and the absence of treatment gaps of greater than 7 days.

Exception to evidence: N/A

Questions for the Committee:

- *Does reducing relapses in OUD lead to better health outcomes?*
- *How strong is the evidence that use of the various medications is associated with reductions in relapse?*

Guidance from the Evidence Algorithm

Process measure supported by systematic review and grading (Box 3) → QQC provided (Box 4) → evidence varied by medication type, but overall: Quantity: high; Quality: moderate; Consistency: moderate (Box 5b) →Moderate

The highest possible rating is 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.

- Performance results for 2010-2015 (for two-year rolling periods) are calculated from commercial pharmacy and medical claims obtained from the Truven MarketScan® Commercial Database. Data were limited to members with at least 6 months of continuous enrollment. The total number of episodes included in the analysis ranged from 17,229 in the 2010-2011 period to 40,379 in the 2014-2015 period.

Summary statistics of performance across states

Time Period	N	Mean	Min	Max	STD	IQR	P10	P25	Median	P75	P90
2010-2011	44	0.250	0.034	0.542	0.086	0.112	0.169	0.188	0.231	0.300	0.333
2011-2012	47	0.246	0.115	0.378	0.066	0.105	0.169	0.189	0.242	0.294	0.342
2012-2013	47	0.286	0.152	0.434	0.072	0.116	0.189	0.229	0.283	0.345	0.387
2013-2014	46	0.287	0.149	0.455	0.076	0.114	0.199	0.235	0.280	0.349	0.394
2014-2015	46	0.307	0.195	0.505	0.071	0.093	0.218	0.256	0.308	0.348	0.395

Summary statistics of performance across commercial health plans

Time Period	N	Mean	Min	Max	STD	IQR	P10	P25	Median	P75	P90
2010-2011	88	0.225	0.034	0.571	0.109	0.140	0.100	0.152	0.198	0.292	0.396
2011-2012	201	0.208	0.033	0.500	0.088	0.112	0.100	0.145	0.200	0.258	0.333
2012-2013	279	0.245	0.000	0.550	0.105	0.143	0.122	0.167	0.238	0.310	0.382
2013-2014	290	0.254	0.045	0.600	0.095	0.129	0.138	0.187	0.241	0.316	0.381
2014-2015	264	0.277	0.025	0.652	0.103	0.137	0.162	0.202	0.263	0.339	0.409

Disparities

- The developer provided overall performance results (not aggregated by state or health plan) by both age group and sex. Performance (i.e., percentage of OUD patients without treatment gap) was lower for both the younger and older age groups as compared to the middle ranges and was lower for women compared to men. It is not clear whether or not the differences between the groups were statistically significant.

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 for other patient subpopulations?*

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 to Support Measure Focus

Comments:

**Table 7 provides evidence of the relationship between the use of mAT for OUD and reduced mortality. Could this measure be considered an outcome measure rather than just a process measure?
The measure does not address counseling as a component of Tx, although there is evidence supporting the combination of MAT plus counseling to be more effective that MAT alone. The measure also does not address the larger issue of underutilization of MAT for OUD.
**Reasonable evidence to support measure focus.
**Evidence supports focus with the exception of limiting the population to people under 65. I see no good reason for this and if we eliminate the age restriction we could better harmonize with NQF 1664
** The relationship between the intermediate outcomes of reduction in relapse and better health and between adherence to effective treatment and reduced relapse seem sound. I'd like to hear discussion from the committee about the pros and cons of considering drug therapy separately from talk therapy.

1b. Performance Gap

Comments:

**There is a significant performance gap. Disparities are questionable.
**Clearly documented.
**There certainly seem to be gaps in care as well as a trend toward improvement. There are also clearly variations around the fairly low general performance. There was some limited stratification (e.g., by sex and age) that seems to show some variation in performance.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Administrative claims data, including pharmacy claims

Specifications:

- This measure is specified for the health plan and regional/state levels of analysis in the behavioral health outpatient setting and clinician office/clinic settings. A higher score indicates better quality.
- Patients included in the measure denominator include those ages 18-64 with a diagnosis of OUD who are continuously enrolled in a commercial health plan and have at least one pharmacy claim for at least one OUD medication (Buprenorphine; Buprenorphine and Naloxone; Naltrexone [oral/ extended-release injectable]; Methadone administration).
- Patients included in the measure numerator include those with 1) at least 180 days from Day 1 of the first OUD medication claim through the measure end date (date the supply from last claim is exhausted, death date, or Dec 31 of year 2 of the measurement period, whichever comes first) and 2) no treatment gaps of more than 7 days (covered days are summed across pharmacy claims based on fill date and days' supply).
- The measure appears to be limited to states or health plans with > 20 patients in the denominator.
- Codes (ICD-9-CM, ICD-10-CM, NDC, HCPCS) and descriptions for the measure data elements are provided, either in the submission form itself or in the supplementary materials provided with the submission.
- No exclusions are defined for the measure (note that members who are not continuously for at least 6 months after the first OUD medication fill during the measurement period are not included in the measure).
    - Methadone can only be legally dispensed as OUD pharmacotherapy in licensed treatment centers, and so is not included in the NDC code list.
    - Buprenorphine can be dispensed either through a pharmacy or in an office/treatment center, and so is identified based on either NDC or HCPCS code.
- The developer suggests stratification of results according to age group, sex, state, and health plan. This stratification is not meant for risk-adjustment purposes.
- A detailed calculation algorithm is provided.
- This measure is not risk-adjusted.

Questions for the Committee:

- *The care settings indicated in the "care setting" and "calculation algorithm" sections of the submission form are inconsistent. Should this measure apply to other settings of care, as indicated in the calculation algorithm?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☐ **Yes** ☐ **No**

Method(s) of reliability testing

- Data used for testing included 2013-2014 commercial pharmacy and medical claims obtained from the Truven MarketScan® Commercial Database. Data were limited to members with at least 6 months of continuous enrollment (n=43,812). A total of 46 states and 179 "pseudo" health plans were included in the analysis (the

Truven database includes information from self-insured employers but does not include a plan ID and therefore health plan membership was imputed based on industry type and MSA).

- Developers conducted a signal-to-noise analysis which is an appropriate method for testing reliability. Specifically, they used the robust Prasad-Rao estimator to estimate the signal and the standard binomial distribution inference to estimate the noise. The calculations were based on the ***average denominator size*** for states (n=471) and health plans (n=33); see Table 1 for information regarding the distribution of patients across states and health plans. The analysis was limited to those health plans with at least 20 members who were eligible for the measure denominator.
- A signal-to-noise analysis quantifies the amount of variation in a performance measure that is due to true differences between providers (i.e., signal) as opposed to measurement error (i.e., noise). Results will vary based on the amount of variation between health plans (or states) and the number of patients treated by each health plan (or state). A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in health plan (or state) performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Results of reliability testing

- State: reliability=0.977; standard deviation=0.008
- Health plans: reliability=0.891; standard deviation=0.040

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
-
- *Reliability was estimated based on the average patient count (928 for states and 52 for health plans). Typically, reliability is lower when the patient count is lower. Is it reasonable to assume that states and health plans will have enough patients eligible for the measure to ensure adequate reliability? If not, is there any data that would indicate how many states or health plans would not have sufficient numbers of patients?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing with measure as specified (Box 2) → Score-level testing (Box 4) → Appropriate method (Box 5) → Moderate certainty that measure results are reliable (Box 6b) → Moderate

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

- The evidence for Naltrexone is weaker than for other medications included in the measure. Evidence for extended-release injectable Naltrexone is based on one randomized trial of 250 participants in Russia; the authors state that the results may not be generalizable. Evidence for oral Naltrexone and the

VA/DoD guideline concludes: "There is insufficient evidence to recommend for or against oral naltrexone for treatment of opioid use disorder."
- Developer bases 7-day gap on the fact that evidence shows "the mortality risk is highest in the first four weeks out of treatment, with many studies showing an increase in mortality in days 1-14 after treatment cessation."

Question for the Committee:
- o *Is it reasonable to include naltrexone in the measure, given the level of evidence?*
- o *Is it reasonable to use the 7-day gap interval for measurement?*

2b2. Validity testing

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

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
- Face validity was assessed by a 10 clinicians with expertise in treating OUD. These individuals were asked to rate their agreement, on a 5-point scale, with the following statement: "*Performance scores resulting from the measure as defined can be used to distinguish good from poor quality*."
- According to NQF guidance, the 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 testing conducted by the developer conforms to NQF's requirements for face validity.

Validity testing results:
- Of the 10 clinicians surveyed, 1 strongly agreed and 7 agreed that results from the measure can be used to distinguish good from poor quality. The remaining 2 clinicians neither agreed nor disagreed.

Questions for the Committee:
- o *Did the clinicians included in the face validity assessment have the appropriate expertise to judge the face validity of the measure?*
- o *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- o *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:
- No exclusions are defined for the measure (note that members who are not continuously covered for at least 6 months after the first AUD medication fill during the measurement period are not included in the measure).

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Questions for the Committee:

o *Process measures generally are not risk adjusted. Do you agree with the developer's decision not to risk-adjust this measure?*

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

- To assess whether differences in performance between states and health plans are meaningful, developers constructed 95% confidence intervals around each state and health plan's performance rate and compared these to the overall state and health plan average performance rate, respectively. They considered the state or health plan rate to be statistically different from the average if the overall mean did not overlap the individual state/health plan confidence interval. The developer used the 2013-2014 Truven MarketScan® Commercial Database for the analysis.
- States
  - 14 of the 46 states (30.4 percent) had scores statistically significantly lower than the state-level mean
  - 15 of the 46 states (32.6 percent) had scores statistically significantly higher than the state-level mean
- Health Plans
  - 49 of 290 health plans (16.9 percent) had scores statistically significantly lower than the health plan-level mean
  - 8 of 179 health plans (2.8 percent) had scores statistically significantly higher than the health plan-level mean

Question for the Committee:

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

2b6. Comparability of data sources/methods: N/A

2b7. Missing Data

- The developers assessed the frequency of missing, zero, or negative values for the "days supply" data element. They report that 687 (0.8%) of the 86,947 individuals eligible for the measure had one or more pharmacy claims with an invalid value for this variable. They interpret this to mean that missing or invalid data for this data element would not substantially impact the measure results.

Guidance from the Validity Algorithm

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

Specifications somewhat consistent with evidence (Box 1) → potential threats to validity assessed (Box2) → empirical validity testing not conducted (Box 3) → face validity systematically assessed (Box 4) → results indicate moderate agreement that the measure results can be used to distinguish quality (Box 5) →Moderate

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

Comments:

**It is unclear what is meant by "outpatient" setting. Is this limited to BH specialty care provider settings? Does the numerator and denominator include patients seen in primary care settings? The majority of MAT for OUD is provided in primary care settings. See AHRQ Technical Brief Dec 6 2016 Medication -Assisted Treatment Models of Care for Opioid Use Disorders in Primary Care Settings. Many State Medicaid data systems still do not offer integrated care and do not have integrated BH data sources which could make capturing the MAT continuity of care challenging.
**No exclusions is concerning. How many dropped out and why.
**OK
**In general the specifications seem reasonable. I agree that the care settings should be clarified. The 20 patients in the denominator seems reasonable for generating interpretable data

2a2. Reliability Testing
Comments:
**Some health plans did not have large numbers.
**OK
**Seems reasonable. Specifically, the n>20 requirement discussed above probably deals with the issue of whether there are sufficient numbers of patients in all states/health plans. Because this is a claims based measure already tested in dozens of plans/states, it seems plausible for widespread implementation

2b1. Validity Specifications
Comments:
**There needs to be sufficient numbers to be valid.
**7 day gap is reasonable. Consider excluding injectable extended release naltrexone
**Seems reasonable.

2b2. Validity Testing
Comments:
**Face validity is not strong.
**Good validity.
**I'd like to hear discussion from treatment experts about whether all of these meds are appropriate and whether a 7 d gap in treatment makes sense as a failure.
**I'd like to hear from the developer what the objections were among the expert panel for the small number of experts that objected to validity or usability.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**Small percentage had invalid numbers.
**good face validity
**Seems likely to show meaningful differences

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 required clinical data elements (e.g., diagnosis) are routinely generated and used during care delivery
- The required data elements are available in electronic sources (i.e., administrative medical and pharmacy claims)
- The developers calculated measure results using a publicly-available claims database. They did not identify any feasibility or implementation issues.

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes

3b. Electronic Sources
3c. Data Collection Strategy
Comments:
**The availability of integrated BH data in state Medicaid data systems will make capturing this data on continuity of MAT for OUD challenging.
**The required data is missing the reason the patient did not continue the medication.
**Seems feasible.
**Because this measure is based on claims, it seems feasible

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details

- The developers suggest this measure is appropriate for use in the CMS Medicaid Adult Core Set. They plan to explore avenues for recommending this measure for use in this program (i.e., potentially through the annual rule-making process)

Improvement results N/A

Unexpected findings (positive or negative) during implementation New measure – none reported.

Potential harms None reported

Vetting of the measure None reported

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?*
- o *Would inclusion of this measure in the Medicaid Adult Core Set be reasonable? Are there other programs that might benefit from inclusion of this measure?*

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

Committee pre-evaluation comments
Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement

4c. Unintended Consequences
Comments:
**Acceptable.
**No comment.
**It is not currently in use. It seems usable if the issues above are addressed.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0004: Initiation and Engagement of Alcohol and Other Drug Dependence Treatment (IET)
- 1664: SUB-3 Alcohol & Other Drug Use Disorder Treatment Provided or Offered at Discharge and SUB-3a Alcohol & Other Drug Use Disorder Treatment at Discharge

Harmonization

- Measure #0004 was discussed with the Behavioral Health Standing Committee in October, 2016. Since that time, the developer has continued its internal re-evaluation of the measure and may provide updates (if any are available at this time).
- Measure #1664 is a facility-level measure for the hospital setting. Differences in denominator definitions (i.e., including individuals with either alcohol or drug use disorder; different age ranges) will be discussed.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because empirical validity testing for the measure score has not been conducted and it has not been vetted by those being measured or others.

Pre-meeting public and member comments

- None received.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** NQF 3175

Measure Title: Continuity of Pharmacotherapy for Opioid Use 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: 1/12/2017

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

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

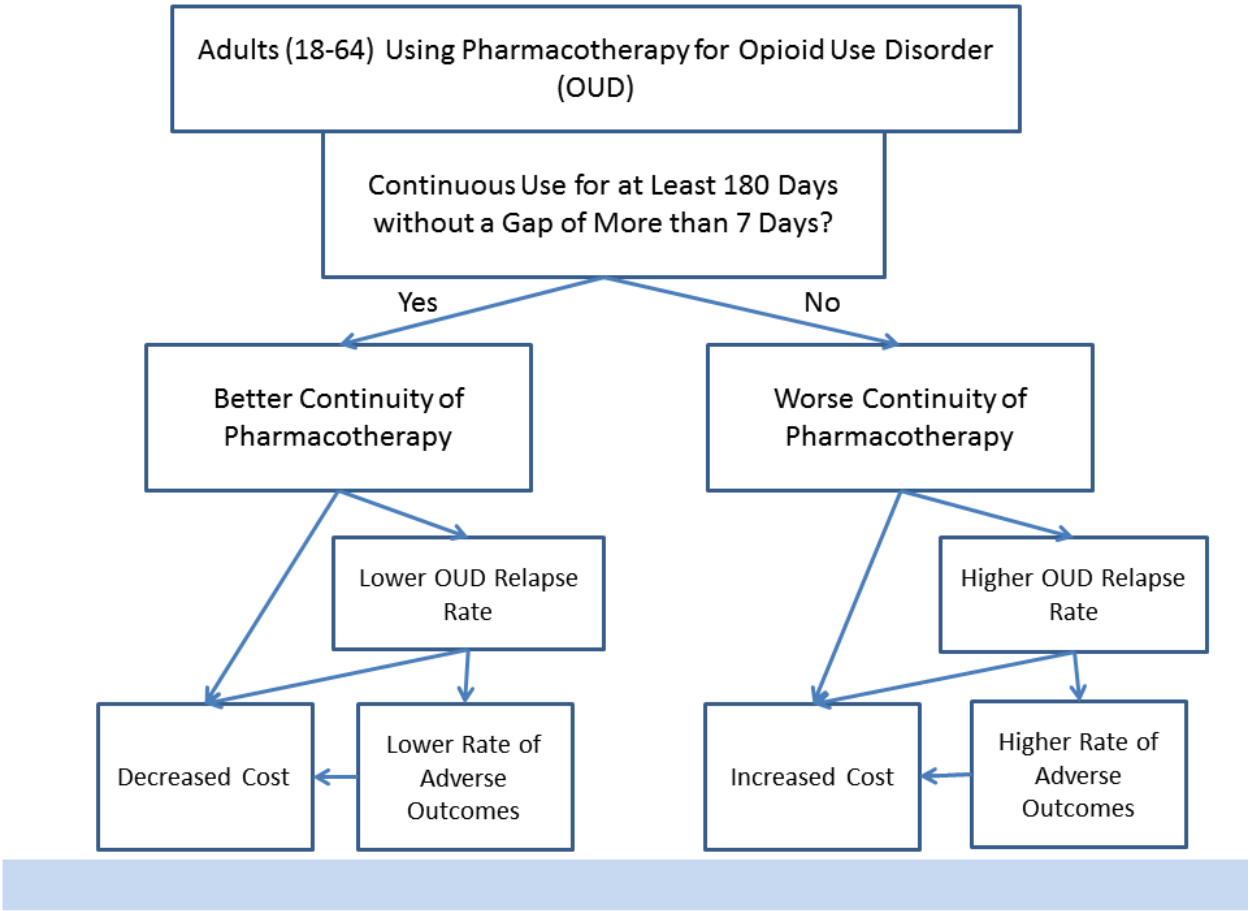
☒ Process: Continuity of Pharmacotherapy for Opioid Use Disorder

☒ Appropriate use measure: Continuous pharmacotherapy for opioid use disorder without a gap of more than 7 days

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

1a.12 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.



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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

The proposed measure is not an outcome measure.

1a.3. SYSTEMATIC REVIEW (SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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: See the sections on "Mortality Risk" and "Justification of Measure Definition" in Section 1a.4 at the end of this file.

On the following pages, we present six tables (Exhibits 1-6) to summarize the evidence cited by the VA/DoD 2015 Guideline on Management of Substance Use Disorders to support the recommendations related to treatment of opioid use disorder with pharmacotherapy. The tables contain paraphrased and verbatim excerpts from the systematic reviews, meta-analyses, and other studies, and are organized by type of medication:

- Exhibit 1: Buprenorphine (meta-analysis; Fareed et al., 2012)
- Exhibit 2: Buprenorphine (systematic review; Mattick et al., 2014)
- Exhibit 3: Methadone (meta-analysis; Bao et al., 2009)
- Exhibit 4: Methadone (systematic review; Mattick et al., 2009)
- Exhibit 5: Extended-Release Injectable Naltrexone (randomized controlled trial; Krupitsky et al., 2011)
- Exhibit 6: Oral Naltrexone (systematic review; Minozzi et al., 2011)

In Section 1a.4 (Other Source of Evidence) at the end of this file, we include two other sources of evidence, under the headings, “Mortality Risk” and “Justification of Measure Definition”.

- **Mortality Risk**: Under this heading, we include a table, Exhibit 7, which summarizes findings from multiple studies that report on the mortality risk during transition of care phases (i.e., treatment initiation and cessation) for use of pharmacotherapy for OUD.
- **Justification of Measure Definition**: Under this heading, we include a summary of evidence that supports the measure definition. This material also appears in Section S.5. (Numerator Details) of the Measure Information Form (MIF).

Exhibit 1. Meta-Analysis Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Buprenorphine for Opioid Use Disorder (Fareed et al., 2012)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Meta-analysis: Fareed A, Vayalapalli S, Casarella J, Drexler K. Effect of buprenorphine dose on treatment outcome. J Addict Dis. 2012;31(1):8-18. Cited in support of Recommendation 8 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 8 (page 38) from 2015 VA/DoD Guideline** 8. For patients with opioid use disorder, we recommend offering one of the following medications considering patient preferences: • Buprenorphine/naloxone • Methadone in an Opioid Treatment Program **(Strong For | Reviewed, New-replaced)**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The meta-analysis did not grade the evidence.
Provide all other grades and definitions from the evidence grading system	The meta-analysis did not grade the evidence.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 8 was “Strong For.” “A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes.” (page 11) “Using these elements, the grade of each recommendation is presented as part of a continuum: “Strong For (or “We recommend offering this option …”) “ (page 11)
Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation: • Balance of desirable and undesirable outcomes

	• Confidence in the quality of the evidence • Patient or provider values and preferences • Other implications, as appropriate, e.g., Resource use, Equity, Acceptability, Feasibility, and Subgroup considerations. Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation. They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies included: Meta-analysis of 21 randomized, controlled, or double-blind clinical trials. A total of 2,703 participants were included in those studies. Quality of studies included: A scale designed to rate the quality of randomized clinical trials (Jadad et al., 1996) was used "to evaluate the quality of the selected studies. The scale is widely used to assess the quality of clinical trials. It is a five-question scale to assess randomization, blindness, and description of participant withdrawals. Studies with a score of 3 or more were included in the meta-analysis" (Fareed et al., 2012). Citations: Fareed A, Vayalapalli S, Casarella J, Drexler K. Effect of buprenorphine dose on treatment outcome. J Addict Dis. 2012;31(1):8-18. Jadad AR; Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996; 17(1):1–12. doi:10.1016/0197-2456(95)00134-4.
Estimates of benefit and consistency across studies	Higher buprenorphine doses of 16-32 mg per day predicted better retention in treatment compared with the lower doses of less than 16 mg per day (univariate analysis: 69% v. 51%; p=0.006). Buprenorphine dose was a significant predictor for retention status in an analysis of predictor variables (univariate analysis: $p<0.00001$; R^2 =0.44). In a multivariate analysis, buprenorphine dose showed significant positive correlation with retention in treatment (p=0.009, adjusted R^2 =0.40). Higher buprenorphine dose predicted less illicit opioid use compared with lower dose (p=0.0019; R^2 =0.29). In a multivariate analysis, buprenorphine dose did not show a significant correlation with illicit opioid use. "Strong evidence exists based on 21 randomized clinical trials that the higher buprenorphine dose may improve retention in buprenorphine maintenance treatment."

What harms were identified?	No harms were identified in the meta-analysis article.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new studies were identified.

Exhibit 2. Systematic Review Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Buprenorphine for Opioid Use Disorder (Mattick et al., 2014)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Systematic review: Mattick RP, Breen C, Kimber J, Davoli M. Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. Cochrane Database Syst Rev. 2014;2:Cd002207. Cited in support of Recommendation 8 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 8 (page 38) from 2015 VA/DoD Guideline** 8. For patients with opioid use disorder, we recommend offering one of the following medications considering patient preferences: • Buprenorphine/naloxone • Methadone in an Opioid Treatment Program **(Strong For | Reviewed, New-replaced)**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The systematic review graded the evidence as follows: Fixed-dose buprenorphine maintenance vs. placebo medication: • Retention in treatment: high quality • Illicit opioid use measured by urinalysis: moderate quality Fixed-dose buprenorphine maintenance vs. fixed-dose methadone maintenance: • Retention in treatment: high quality • Illicit opioid use measured by urinalysis: moderate quality **GRADE Working Group Grades of Evidence** **High quality:** Further research is very unlikely to change our confidence in the estimate of effect. **Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Provide all other grades and definitions from the evidence grading system	**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. **Very low quality:** We are very uncertain about the estimate.

Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 8 was "Strong For." "A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes." (page 11) "Using these elements, the grade of each recommendation is presented as part of a continuum: "Strong For (or "We recommend offering this option …") " (page 11)
Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation: • Balance of desirable and undesirable outcomes • Confidence in the quality of the evidence • Patient or provider values and preferences • Other implications, as appropriate, e.g., Resource use, Equity, Acceptability, Feasibility, and Subgroup considerations. Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation. They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: The results are based on 5430 patients in 31 RCTs. Quality of studies: "The clinical trials represented in this review are of reasonable quality, and whilst many of them did not fully explain how randomization was concealed, they appear to have used doses which are clinically relevant and to have treated participants for significant periods of time. Moreover, despite the tendency of randomised studies to include selected populations, characteristics of drug users enrolled in the studies included in this review appear to be heterogeneous enough to allow generalisability of the results across different clinical and cultural settings. Based on the nature of the trials, it would appear the external validity or generalisability of the results is quite good, particularly from those trials which have used large sample sizes and adequate doses." "Thirteen studies provided an adequate sequence generation for the randomisation process." "Of the 31 studies included in this review, 22 were reportedly conducted under double-blind conditions." "All the studies have been judged to be at low risk of bias because all used the intention-to-treat (ITT) principle."

Estimates of benefit and consistency across studies	The results are based on 5430 patients in 31 RCTs. **Fixed-dose studies of buprenorphine vs. placebo:** "There is high quality of evidence that buprenorphine was superior to placebo medication in retention of participants in treatment at all doses examined. Specifically, buprenorphine retained participants better than placebo: at low doses (2 - 6 mg), 5 studies, 1131 participants, risk ratio (RR) 1.50; 95% confidence interval (CI) 1.19 to 1.88; at medium doses (7 - 15 mg), 4 studies, 887 participants, RR 1.74; 95% CI 1.06 to 2.87; and at high doses (≥ 16 mg), 5 studies, 1001 participants, RR 1.82; 95% CI 1.15 to 2.90. However, there is moderate quality of evidence that only high-dose buprenorphine (≥ 16 mg) was more effective than placebo in suppressing illicit opioid use measured by urinanalysis in the trials, 3 studies, 729 participants, standardised mean difference (SMD) -1.17; 95% CI -1.85 to -0.49, notably, low-dose, (2 studies, 487 participants, SMD 0.10; 95% CI -0.80 to 1.01), and medium-dose, (2 studies, 463 participants, SMD -0.08; 95% CI -0.78 to 0.62) buprenorphine did not suppress illicit opioid use measured by urinanalysis better than placebo." **Flexible-dose studies of buprenorphine vs. methadone:** "There is high quality of evidence that buprenorphine in flexible doses adjusted to participant need,was less effective than methadone in retaining participants, 5 studies, 788 participants, RR 0.83; 95% CI 0.72 to 0.95. For those retained in treatment, no difference was observed in suppression of opioid use as measured by urinalysis, 8 studies, 1027 participants, SMD -0.11; 95% CI -0.23 to 0.02 or self report, 4 studies, 501 participants, SMD -0.11; 95% CI -0.28 to 0.07, with moderate quality of evidence." **Fixed-dose studies of buprenorphine vs. methadone:** "Consistent with the results in the flexible-dose studies, in low fixed-dose studies, methadone (≤ 40 mg) was more likely to retain participants than low-dose buprenorphine (2 - 6 mg), (3 studies, 253 participants, RR 0.67; 95% CI: 0.52 to 0.87). However, we found contrary results at medium dose and high dose: there was no difference between medium-dose buprenorphine (7 - 15 mg) and medium-dose methadone (40 - 85 mg) in retention, (7 studies, 780 participants, RR 0.87; 95% CI 0.69 to 1.10) or in suppression of illicit opioid use as measured by urines, (4 studies, 476 participants, SMD0.25; 95%CI -0.08 to 0.58) or self-report of illicit opioid use, (2 studies, 174 participants, SMD -0.82; 95% CI -1.83 to 0.19). Similarly, there was no difference between high-dose buprenorphine (≥ 16 mg) and high-dose methadone (≥ 85 mg) in retention (RR 0.79; 95% CI 0.20 to 3.16) or suppression of self-reported heroin use (SMD -0.73; 95% CI -1.08 to -0.37) (1 study, 134 participants)." RR=relative risk; SMD= standardised mean difference
What harms were identified?	Ten studies collected data on a wide range of adverse events or side effects, but not all reported them. "[T]wo studies compared adverse events statistically, finding no difference between methadone and buprenorphine, except for a single result indicating more sedation among those using methadone." Adverse events reported by the two studies included, but were not limited to: sedation, insomnia, headache, depression, sweating, and dyspepsia.
Identify any new studies conducted	No new studies were identified.

since the SR. Do the new studies change the conclusions from the SR?	

Exhibit 3. Meta-Analysis Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Methadone for Opioid Use Disorder (Bao et al., 2009)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Meta-analysis: Bao YP, Liu ZM, Epstein DH, Du C, Shi J, Lu L. A meta-analysis of retention in methadone maintenance by dose and dosing strategy. Am J Drug Alcohol Abuse. 2009;35(1):28-33. Cited in support of Recommendation 8 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 8 (page 38) from 2015 VA/DoD Guideline** 8. For patients with opioid use disorder, we recommend offering one of the following medications considering patient preferences: • Buprenorphine/naloxone • Methadone in an Opioid Treatment Program **(Strong For | Reviewed, New-replaced)**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The meta-analysis did not grade the evidence.
Provide all other grades and definitions from the evidence grading system	The meta-analysis did not grade the evidence.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 8 was “Strong For.” “A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes.” (page 11) “Using these elements, the grade of each recommendation is presented as part of a continuum: “Strong For (or “We recommend offering this option …”) “ (page 11)
Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation: • Balance of desirable and undesirable outcomes

	• Confidence in the quality of the evidence • Patient or provider values and preferences • Other implications, as appropriate, e.g., Resource use, Equity, Acceptability, Feasibility, and Subgroup considerations. Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation. They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: 18 studies with 2831 participants. Quality of studies: All 18 studies were randomized, controlled, double-blind clinical trials with methadone maintenance treatment (MMT) as at least one of the treatments. These study design characteristics were part of the criteria for selecting studies for inclusion.
Estimates of benefit and consistency across studies	In univariate analyses, doses of MMT greater than or equal to 60 mg/day were associated with greater retention than doses less than 60 mg/day at 3-6 months (62.5% vs. 50.6%; p=0.0005) and 6-12 months (57.0% vs. 42.5%; p<0.0001). Flexible dosing was associated with greater retention than fixed dosing strategies at 3-6 months (61.0% vs. 49.9%; p=0.0007) and 6-12 months (61.7% vs. 45.9%; p<0.0001). In multilevel analyses (follow-up duration, dose, and dosing strategy), retention was greater with methadone doses ≥ 60 mg/day than with doses <60 mg/day (OR: 1.74, 95% CI: 1.43–2.11). Similarly, retention was greater with flexible-dose strategies than with fixed-dose strategies (OR: 1.72, 95% CI: 1.41–2.11). OR=odds ratio
What harms were identified?	Adverse events or harms were not identified in the systematic review.
Identify any new studies conducted since the SR. Do the new studies change the	No new studies were identified.

conclusions from the SR?	

Exhibit 4. Systematic Review Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Methadone for Opioid Use Disorder (Mattick et al., 2009)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Systematic review: Mattick RP, Breen C, Kimber J, Davoli M. Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence. Cochrane Database Syst Rev. 2009(3):Cd002209. Cited in support of Recommendation 8 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 8 (page 38) from 2015 VA/DoD Guideline** 8. For patients with opioid use disorder, we recommend offering one of the following medications considering patient preferences: • Buprenorphine/naloxone • Methadone in an Opioid Treatment Program **(Strong For | Reviewed, New-replaced)**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The systematic review graded the evidence as follows: Methadone maintenance treatment vs. No methadone maintenance treatment: • Retention in treatment -Old studies (pre 2000): high quality • Retention in treatment - New studies: high quality • Morphine positive urine or hair analysis: high quality • Criminal activity: moderate quality **GRADE Working Group Grades of Evidence** **High quality:** Further research is very unlikely to change our confidence in the estimate of effect. **Moderate quality:** Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Provide all other grades and definitions from the evidence grading system	**Low quality:** Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate. **Very low quality:** We are very uncertain about the estimate.
Grade assigned to the **recommendation**	The grade assigned to Recommendation 8 was "Strong For." "A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes." (page 11)

with definition of the grade	"Using these elements, the grade of each recommendation is presented as part of a continuum: "Strong For (or "We recommend offering this option …") " (page 11)
Provide all other grades and definitions from the recommendation grading system	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation: • Balance of desirable and undesirable outcomes • Confidence in the quality of the evidence • Patient or provider values and preferences • Other implications, as appropriate, e.g., Resource use, Equity, Acceptability, Feasibility, and Subgroup considerations. Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation. They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: Eleven randomized clinical trials were included in the review from 1969 to 2008 Quality of studies: "Eleven studies met the criteria for inclusion in this review, all were randomised clinical trials, two were double-blind. … The sequence generation was inadequate in one study, adequate in five studies and unclear in the remaining studies. The allocation of concealment was adequate in three studies and unclear in the remaining studies."
Estimates of benefit and consistency across studies	The results are based on 1969 patients in 11 randomized clinical trials. "Methadone appeared statistically significantly more effective than non-pharmacological approaches in retaining patients in treatment and in the suppression of heroin use as measured by self report and urine/hair analysis (6 RCTs, RR = 0.66; 95% CI 0.56-0.78), but not statistically different in criminal activity (3 RCTs, RR=0.39; 95% CI 0.12-1.25) or mortality (4 RCTs, RR=0.48; 95% CI: 0.10-2.39)." RR=risk ratio
What harms were identified?	No harms were discussed in the systematic review
Identify any new studies conducted	No new studies were identified.

since the SR. Do the new studies change the conclusions from the SR?	

Exhibit 5. Randomized Controlled Trial Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Extended-Release Injectable Naltrexone for Opioid Use Disorder (Krupitsky et al., 2011)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Randomized controlled trial: Krupitsky E, Nunes EV, Ling W, Illeperuma A, Gastfriend DR, Silverman BL. Injectable extended-release naltrexone for opioid dependence: A double-blind, placebo-controlled, multicentre randomised trial. Lancet. Apr 30 2011;377(9776):1506-1513. Cited in support of Recommendation 11 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 11 (page 38) from 2015 VA/DoD Guideline** 11.For patients with opioid use disorder for whom opioid agonist treatment is contraindicated, unacceptable, unavailable, or discontinued and who have established abstinence for a sufficient period of time (see narrative), we recommend offering: ▪ Extended-release injectable naltrexone **(Strong For | Reviewed, New-replaced)**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The Guideline did not assign a grade to the evidence in this study.
Provide all other grades and definitions from the evidence grading system	The Guideline did not assign a grade to the evidence in this study.
Grade assigned to the **recommendation** with definition of the grade	The grade assigned to Recommendation 11 was “Strong For.” “A strong recommendation indicates that the Work Group is highly confident that desirable outcomes outweigh undesirable outcomes.” (page 11) “Using these elements, the grade of each recommendation is presented as part of a continuum: “Strong For (or “We recommend offering this option …”) “ (page 11)
Provide all other grades and definitions from the	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation: • Balance of desirable and undesirable outcomes

recommendation grading system	• Confidence in the quality of the evidence • Patient or provider values and preferences • Other implications, as appropriate, e.g., Resource use, Equity, Acceptability, Feasibility, and Subgroup considerations. Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation. They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity: One trial with 250 participants. Quality of study: A double-blind, placebo-controlled, randomized, 24-week trial of injectable extended-release naltrexone patients with opioid dependence disorder at 13 clinical sites in Russia. The study required "that patients have someone available to supervise attendance, the provision of individual counselling, the absence of alternative treatments (eg, methadone or buprenorphine) in Russia, and the promise of active XR-NTX treatment for all patients after 6 months in the subsequent open-label extension safety study." Therefore, the results may not be generalizable.
Estimates of benefit and consistency across studies	The median proportion of weeks of confirmed abstinence was significantly higher in the naltrexone group than in the placebo group (90.0% for naltrexone vs. 35.0% for placebo; p=0.0002). The proportion of patients with total confirmed abstinence was higher in the naltrexone group than the placebo group (RR=1.58; 95% CI, 1.06 to 2.36; p=0.0224). Comparing clinical outcomes between the naltrexone and placebo groups yielded the following results: proportion of self-reported opioid-free days over the 24 weeks (99.2% for naltrexone vs. 60.4% for placebo; p=0.0004), mean change in opioid craving score from baseline (-10.1 for naltrexone vs. 0.7 for placebo; p<0.0001), number of days of retention (>168 days for naltrexone vs. 96 days for placebo; p=0·0042), and number of participants with positive naloxone challenge test (1 for naltrexone vs. 17 for placebo; p<0.0001).
What harms were identified?	103 (41%) of 250 patients experienced at least one adverse event; a higher proportion of patients in the naltrexone group had at least one adverse event (e.g., nasopharyngitis, insomnia, hypertension, influenza, injection site pain) than in the placebo group (p=0.005).
Identify any new studies conducted	No new studies were identified.

since the SR. Do the new studies change the conclusions from the SR?	

Exhibit 6. Systematic Review Cited by the Department of Veteran Affairs, Department of Defense (VA/DoD) Guideline on Management of Substance Use Disorders: Oral Naltrexone for Opioid Use Disorder (Minozzi et al., 2011)

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Systematic review: Minozzi S, Amato L, Vecchi S, Davoli M, Kirchmayer U, Verster A. Oral naltrexone maintenance treatment for opioid dependence. Cochrane Database Syst Rev. 2011(4):Cd001333. Cited in support of Recommendation 12 (see below) by: Department of Veteran Affairs, Department of Defense (VA/DoD). (2015). VA/DoD clinical practice guideline for the management of substance use disorders. Version 3.0. Washington (DC): Department of Veteran Affairs, Department of Defense; 2015 December. Available at http://www.healthquality.va.gov/guidelines/MH/sud/VADoDSUDCPGRevised22216.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.	**Recommendation 12 (page 38) from 2015 VA/DoD Guideline** 12. There is insufficient evidence to recommend for or against oral naltrexone for treatment of opioid use disorder. **(N/A | Reviewed, New-replaced)**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The systematic review did not grade the evidence
Provide all other grades and definitions from the evidence grading system	The systematic review did not grade the evidence
Grade assigned to the **recommendation** with definition of the grade	A grade was not assigned to Recommendation 12 because of insufficient evidence to recommend for or against oral naltrexone for treatment of opioid use disorder (page 38).
Provide all other grades and definitions from the	The [DoD/VA] Work Group used the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system to assess the quality of the evidence base and assign a grade for the strength for each recommendation. The GRADE system uses the following four domains to assess the strength of each recommendation: • Balance of desirable and undesirable outcomes

recommendation grading system	• Confidence in the quality of the evidence • Patient or provider values and preferences • Other implications, as appropriate, e.g., Resource use, Equity, Acceptability, Feasibility, and Subgroup considerations. Using this system, the [DoD/VA] Work Group determined the relative strength of each recommendation (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 give a weak recommendation. They also determined the direction of each recommendation (For or Against). 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 …")
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Quantity of studies: Systematic review of 13 studies with 1158 patients. Quality of studies: "The majority of the studies [were] not of high quality. Only two studies reported information about sequence generation and only three about allocation concealment. Eight out of thirteen studies were double blind, the other were open trial. Nevertheless we think that this did not introduce bias in the main outcomes addressed in this review, because the retention in treatment is an objective measure and abstinence is assessed by urine analysis in all trials. Incomplete outcome data was addressed correctly in the majority of the studies and in any case it does not introduce bias for the outcome retention and retention and abstinence which are the main outcomes on which the review is focused." "Moreover it should be noted that most of the comparisons were underpowered, with few studies and participants included in the analyses, …limit[ing] the strength of the evidence as well as the completeness and the applicability." "The main problem associated with oral naltrexone was the high dropout rate: 72% in our included studies; to overcome this problem the sustained released naltrexone has been proposed and is currently being assessed. However, a parallel Cochrane review on sustained released naltrexone (Lobmaier 2008) concluded that there is insufficient evidence to evaluate the effectiveness of sustained release naltrexone for treatment of opioid dependence. A concern about lack of information on mortality data, limit the applicability of the produced evidence due to the relevant problem of fatal overdoses in naltrexone treated patients."
Estimates of benefit and	"Comparing naltrexone versus placebo or no pharmacological treatments, no statistically significant difference was noted for all the primary outcomes considered.

consistency across studies	The only outcome statistically significant in favour of naltrexone is reincarceration, RR 0.47 (95% CI 0.26-0.84), but results come only from two studies. Considering only studies where patients were forced to adherence a statistical significant difference in favour of naltrexone was found for retention and abstinence, RR 2.93 (95%CI 1.66-5.18)." RR=risk ratio Naltrexone versus placebo or no pharmacological treatment 1.1 Retention in treatment: two studies; 88 participants; RR 1.18 (95% CI 0.72-1.91); no statistically significant difference 1.2 Retention and abstinence: six studies; 393 participants; RR 1.43 (95% CI 0.72-2.82); no statistically significant difference 1.3 Abstinence: four studies; 143 participants; RR 1.39 (95% CI 0.61-3.17); no statistically significant difference 1.4 Abstinence at follow up: three studies; 116 participants; RR 1.28 (95% CI 0.80-2.08); no statistically significant difference 1.5 Side effects: four studies; 159 participants; RR 1.29 (95% CI 0.54-3.11); no statistically significant difference 1.6 Reincarceration: two studies; 86 participants; RR 0.47 (95% CI 0.26-0.84); results in favor of naltrexone
What harms were identified?	Four of the 13 studies (159 participants) reported side effects; these varied by study but included abdominal discomfort, sleep disturbances, loss of appetite, diarrhea, and nausea. However, there was no statistically significant difference between treatment and control groups (RR 1.29; 95% CI 0.54-3.11).
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	No new studies were identified.

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.

In this section, we provide two supplementary sources of evidence on which we are basing the measure.

Mortality Risk: The first source summarizes findings from our review of studies that looked at the mortality risk during transition of care phases for OUD pharmacotherapy (treatment initiation and cessation). This evidence supports the recommendation for no gaps in care of more than 7 days.Each of the sections related to this topic (presented below) are labeled with the header, "Mortality Risk".

Justification of Measure Definition: The second source provides evidence that justifies the measure definition. This evidence also appears in the Numerator Details section of the MIF (see Section S.5). Each of the sections related to this topic (presented below) are labeled with the header, "Justification of Measure Definition".

Mortality Risk

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

Exhibit 7 summarizes findings from our review of studies that looked at the mortality risk during transition of care phases for OUD pharmacotherapy (treatment initiation and cessation). The vast majority of studies found evidence for increased mortality during those periods and the results were consistent for the different MAT drugs. Across all the medications, the mortality risk is highest in the first four weeks out of treatment, with many studies showing an increase in mortality in days 1-14 after treatment cessation. This evidence supports the recommendation for no gaps in care of more than 7 days.

Exhibit 7. Mortality Subsequent to Pharmacotherapy Initiation and Cessation.

Author (Year)	Drug(s)	Increased Mortality Risk	Timeframe (after treatment initiation or cessation)
Cornish et al., 2010 Degenhardt et al., 2009	Buprenorphine	No	After treatment initiation
Cornish et al., 2010 Degenhardt et al., 2009	Buprenorphine	Yes	After treatment cessation
Cousins et al., 2016 Caplehorn, 1998 Caplehorn and Drummer, 1999 Cornish et al., 2010 Cousins et al., 2011 Degenhardt et al., 2009 Gibson et al., 2007 Tait et al., 2008 Zador et al., 2000	Methadone	No	After treatment initiation
Cornish et al., 2010 Cousins et al., 2016 Davoli et al., 2007 Degenhardt et al., 2009	Methadone	Yes	After treatment initiation After treatment cessation
Pierce et al., 2016	Methadone/ Buprenorphine	No	After treatment initiation
Cornish et al., 2010 Cornish et al., 2010 Pierce et al., 2016	Methadone/ Buprenorphine	Yes	After treatment initiation After treatment cessation
Tait et al., 2008 Gibson & Degenhardt , 2007	Naltrexone	No	After treatment initiation
Tait et al., 2008 Gibson & Degenhardt , 2007	Naltrexone	Yes	After treatment cessation

Mortality Risk

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

We conducted a targeted literature search supplemented by a manual search of the references cited in relevant articles.

Mortality Risk

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

Caplehorn JRM. Deaths in the first two weeks of maintence treatment in NSW in 1994: Identifying cases of iatrogenic methadone toxicity.Drug and Alcohol Review. 1998;17(1):9-17.

Caplehorn JRM, Drummer OH. Mortality associated with New South Wales methadone program in 1994: lives lost and saved. Med J Aust. 1999;170(3):104-109.

Cornish R, Macleod J, Strang J, Vickerman P, Hickman M. Risk of death during and after opiate substitution treatment in primary care: prospective observational study in UK General Practice Research Database. Bmj. 2010;341:c5475.

Cousins G, Teljeur C, Motterlini N, McCowan C, Dimitrov BD, Fahey T. Risk of drug-related mortality during periods of transition in methadone maintenance treatment: a cohort study. J Subst Abuse Treat 2011; 41: 252–60.

Cousins G, Boland F, Courtney B, Barry J, Lyons S, Fahey T. Risk of mortality on and off methadone substitution treatment in primary care: a national cohort study. Addiction. 2016;111(1):73-82.

Davoli M, Bargagli AM, Perucci CA, et al. Risk of fatal overdose during and after specialist drug treatment: the VEdeTTE study, a national multisite prospective cohort study. Addiction. 2007;102:1954-9.

Degenhardt L, Randall D, Hall W, Law M, Butler T, Burns L. Mortality among clients of a state-wide opioid pharmacotherapy program over 20 years: risk factors and lives saved. Drug and alcohol dependence. 2009;105:9-15.

Gibson AE, Degenhardt LJ. Mortality related to pharmacotherapies for opioid dependence: a comparative analysis of coronial records. Drug Alcohol Rev. 2007; 26(4), 405-410.

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, Jones A, et al. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. Addiction. 2016;111:298-308.

Tait RJ, Ngo HT, Hulse GK. Mortality in heroin users 3 years after naltrexone implant or methadone maintenance treatment. J Subst Abuse Treat, 2008;35(2), 116-124.

Weiss RD; Potter JS; Griffin ML, et al. Long-term outcomes from the National Drug Abuse Treatment Clinical Trials Network Prescription Opioid Addiction Treatment Study. Drug and Alcohol Dependence. 2015;150:112-119.

Zador D, Sunjic S. Deaths inmethadone maintenance treatment in New South Wales, Australia 1990-1995. Addiction. 2000;95(1):77-84.

Justification of Measure Definition

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

We define treatment continuity as (1) receiving at least 180 days of treatment and (2) no gaps in medication use of more than 7 days.

Our definition of minimum duration is based on the fact that the FDA registration trials for OUD drugs studied the effect of treatment over three to six months (US FDAa, undated; US FDAb, undated), and we have no evidence for effectiveness of shorter durations. In addition, several recommendations support a minimum six-month treatment period as the risk of relapse is the highest in the first 6-12 months after start of opioid abstinence (US FDAa, undated; US FDAb, undated; US DHHS, 2015). Longer treatment duration is associated with better outcomes compared to shorter treatments and the best outcomes have been observed among patients in long-term methadone maintenance programs ("Effective medical treatment of opiate addiction", 1998; Gruber et al., 2008; Moos et al., 1999; NIDA, 1999; Ouimette et al., 1998; Peles et al., 2013). Studies with long-term follow-up suggest that ongoing pharmacotherapy is associated with improved odds of opioid abstinence (Hser et al., 2015; Weiss et al., 2015). We did not specify a maximum duration of treatment, as no upper limit for duration of treatment has been empirically established (US DHHS, 2015).

We opted for using a treatment gap of more than seven days in our definition, given that the measure includes three active ingredients with different pharmacological profiles. There is substantial evidence for an elevated mortality risk immediately after treatment cessation (Cornish et al., 2010; Cousins et al., 2016; Davoli et al, 2007; Degenhardt et al., 2009; Gibson & Degenhardt, 2007;Pierce et al., 2016). Research suggests that methadone tolerance is lost after three days and this three-day threshold has been used in other observational methadone studies and in developing a United Kingdom treatment guideline which recommends revaluating patients for intoxication and withdrawal after a three-day methadone treatment gap (Cousins et al., 2016; Cousins et al., 2011; "Drug Misuse and Dependence—Guidelines on Clinical Management", 1999). Across all the medications, the mortality risk is highest in the first four weeks out of treatment, with many studies showing an increase in mortality in days 1-14 after treatment cessation.

Justification of Measure Definition

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

We conducted a targeted literature search supplemented by a manual search of the references cited in relevant articles.

Justification of Measure Definition

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

Cornish R, Macleod J, Strang J, Vickerman P, Hickman M. Risk of death during and after opiate substitution treatment in primary care: prospective observational study in UK General Practice Research Database. Bmj. 2010;341:c5475.

Cousins G, Teljeur C, Motterlini N, McCowan C, Dimitrov BD, Fahey T. Risk of drug-related mortality during periods of transition in methadone maintenance treatment: a cohort study. J Subst Abuse Treat 2011; 41: 252–60.

Cousins G, Boland F, Courtney B, Barry J, Lyons S, Fahey T. Risk of mortality on and off methadone substitution treatment in primary care: a national cohort study. Addiction. 2016;111(1):73-82.

Davoli M, Bargagli AM, Perucci CA, et al. Risk of fatal overdose during and after specialist drug treatment: the VEdeTTE study, a national multisite prospective cohort study. Addiction. 2007;102:1954-9.

Degenhardt L, Randall D, Hall W, Law M, Butler T, Burns L. Mortality among clients of a state-wide opioid pharmacotherapy program over 20 years: risk factors and lives saved. Drug and alcohol dependence. 2009;105:9-15.

“Drug Misuse and Dependence—Guidelines on Clinical Management.” Scottish Office Department of Health, Welsh Office, Social Services Northern Ireland. London: Stationery Office, 1999.

Effective medical treatment of opiate addiction. National Consensus Development Panel on Effective Medical Treatment of Opiate Addiction. JAMA.1998;280:1936-1943.

Gibson AE, Degenhardt LJ. Mortality related to pharmacotherapies for opioid dependence: a comparative analysis of coronial records. Drug Alcohol Rev. 2007; 26(4), 405-410.

Gruber VA, Delucchi KL, Kielstein A, Batki SL. A randomized trial of 6-month methadone maintenance with standard or minimal counseling versus 21-day methadone detoxification. Drug and Alcohol Dependence. 2008;94(1-3):199-206.

Hser YI, Evans E, Grella C, Ling W, Anglin D. Long-term course of opioid addiction. Harvard Review of Psychiatry. 2015;23(2):76-89.

Moos RH, Finney JW, Ouimette PC, Suchinsky RT. A comparative evaluation of substance abuse treatment: I. Treatment orientation, amount of care, and 1-year outcomes. Alcohol Clin Exp Res. 1999;23(3):529-36.

National Institute on Drug Abuse (NIDA). Principles of Drug Addiction Treatment: A Research-Based Guide. NIH Publication No. 99–4180. Rockville, MD: NIDA, 1999, reprinted 2000

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522

Peles E, Schreiber S, Adelson M. Opiate-dependent patients on a waiting list for methadone maintenance treatment are at high risk for mortality until treatment entry. J Addict Med. 2013;7(3):177-82..

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, Jones A, et al. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. Addiction. 2016;111:298-308.

U.S. Department of Health and Human Services Assistant Secretary for Planning and Evaluation Office of Disability, Aging and Long-Term Care Policy. Review of Medication-Assisted Treatment Guidelines and Measures for Opioid and Alcohol Use. Washington, DC, 2015. Accessed November 9, 2016 at: https://aspe.hhs.gov/sites/default/files/pdf/205171/MATguidelines.pdf

U.S. Food and Drug Administration (FDA) (a). REVIA Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/018932s017lbl.pdf

U.S. Food and Drug Administration (FDA) (b). VIVITROL Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/021897lbl.pdf

Weiss RD; Potter JS; Griffin ML, et al. Long-term outcomes from the National Drug Abuse Treatment Clinical Trials Network Prescription Opioid Addiction Treatment Study. Drug and Alcohol Dependence. 2015;150:112-119.

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_3175_OUD_Evidence_Form_1-12-17_To_NQF.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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 PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. (Describe how and from whom their input was obtained.)
IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

The rapidly rising number of deaths and near-deaths from opioid overdoses over the past several years has brought the issue of treating opioid use disorder (OUD) to the forefront of the policy agenda. The Surgeon General , who mailed a call to action to 2.3 million doctors, nurses, dentists, and other clinicians asking them to help address this escalating epidemic (Murthy, 2016), and the governors of many states have prioritized improving access to prevention and treatment of OUD.

The high prevalence of OUD has increased the sense of urgency: According to the 2014 National Survey on Drug Use and Health (NSDUH), 1.7 million adults 18 years and older were classified as having a pain reliever use disorder and 886,000 adults had used heroin in the past year (SAMHSA, 2015). In 2014, there were 489,532 episodes for OUD treatment, including outpatient treatment, detoxification, and residential treatment, in the Treatment Episode Data Set (TEDS) (SAMHSA, 2016).

Medication-assisted treatment (i.e., pharmacotherapy combined with counseling) is an evidence-based and effective treatment option for patients with OUD. Individuals with OUD receiving pharmacotherapy have significantly lower rates of mortality while they continue to receive medication, compared to individuals who were not receiving medication (Brugal et al., 2005; Cornish et al., 2010; Cousins et al., 2016; Davoli et al, 2007; Degenhardt et al., 2009; Gibson & Degenhardt, 2007; Pierce et al., 2016). But OUD medications remain markedly underutilized (Volkow et al., 2014). Of the almost half million aforementioned episodes, medication-assisted treatment was planned for only 25.4 percent (20.7 percent in outpatient treatment, 3.6 percent in detoxification, and 1.1 percent in residential treatment) (SAMHSA, 2016).

Additionally, ensuring treatment continuity is critical to the success of medication-assisted treatment. Longer treatment duration for individuals with OUD is associated with better outcomes and the best outcomes have been observed in patients in long-term methadone maintenance programs ("Effective medical treatment of opiate addiction", 1998; Moos et al., 1999; NIDA 1999; Ouimette et al., 1998; Peles et al., 2013).

In addition, there is strong evidence for increased mortality when individuals with OUD transition off pharmacotherapy, both in the short term (within the first 0-4 weeks after discontinuing pharmacotherapy) and over the long term (Cornish et al., 2010; Cousins et al., 2016; Davoli et al, 2007; Degenhardt et al., 2009; Gibson & Degenhardt, 2007; Pierce et al., 2016).

Therefore, the proposed measure focuses on continuity of pharmacotherapy, defined as treatment duration of at least 180 days and absence of treatment gaps of greater than 7 days. Several important benefits related to quality improvement are envisioned with the implementation of this measure. First, the measure will help health plans and providers to identify individuals with OUD who are non-adherent to or discontinue pharmacotherapy. As a result, this measure will encourage health plans and providers to develop communication and education tools and processes to improve treatment continuity in their patients with OUD. Improved treatment continuity is expected to result in lower rates of relapse, and less substance use-related morbidity and mortality. Adoption of this performance measure has the potential to improve quality of care for individuals with OUD and, therefore, advance quality of care by engaging patients as partners in their care, and promoting effective communication and coordination of care, priority areas identified in the National Quality Strategy.

CITATIONS

Brugal MT, Domingo-Salvany A, Puig R, et al. Evaluating the impact of methadone maintenance programmes on mortality due to overdose and aids in a cohort of heroin users in Spain. Addiction. 2005;100(7):981-989.

Cornish R, Macleod J, Strang J, Vickerman P, Hickman M. Risk of death during and after opiate substitution treatment in primary care: prospective observational study in UK General Practice Research Database. BMJ. 2010;341:c5475.

Cousins G, Boland F, Courtney B, Barry J, Lyons S, Fahey T. Risk of mortality on and off methadone substitution treatment in primary care: a national cohort study. Addiction. 2016;111(1):73-82.

Davoli M, Bargagli AM, Perucci CA, et al. Risk of fatal overdose during and after specialist drug treatment: the VEdeTTE study, a national multisite prospective cohort study. Addiction. 2007;102:1954-9.

Degenhardt L, Randall D, Hall W, Law M, Butler T, Burns L. Mortality among clients of a state-wide opioid pharmacotherapy program over 20 years: risk factors and lives saved. Drug and alcohol dependence. 2009;105:9-15.

Effective medical treatment of opiate addiction. National Consensus Development Panel on Effective Medical Treatment of Opiate Addiction. JAMA.1998;280:1936-1943.

Gibson AE, Degenhardt LJ. Mortality related to pharmacotherapies for opioid dependence: a comparative analysis of coronial records. Drug Alcohol Rev. 2007; 26(4), 405-410.

Moos RH, Finney JW, Ouimette PC, Suchinsky RT. A comparative evaluation of substance abuse treatment: I. Treatment orientation, amount of care, and 1-year outcomes. Alcohol Clin Exp Res. 1999;23(3):529-36.

Murthy VH. Ending the Opioid Epidemic - A Call to Action. N Engl J Med. 2016 Dec 22;375(25):2413-2415. doi: 10.1056/NEJMp1612578. Epub 2016 Nov 9.

National Institute on Drug Abuse (NIDA). Principles of Drug Addiction Treatment: A Research-Based Guide. NIH Publication No. 99–4180. Rockville, MD: NIDA, 1999, reprinted 2000.

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522.

Peles E, Schreiber S, Adelson M. Opiate-dependent patients on a waiting list for methadone maintenance treatment are at high risk for mortality until treatment entry. J Addict Med. 2013;7(3):177-82.

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, Jones A, et al. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. Addiction. 2016;111:298-308.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (2015). Behavioral health trends in the United States: results from the 2014 National Survey on Drug Use and Health. Rockville, MD:

Substance Abuse and Mental Health Services Administration, 2015: 7-12 (http://www .samhsa .gov/ data/sites/ default/ files/ NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf).

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. (2016). Treatment Episode Data Set (TEDS): 2004-2014. National Admissions to Substance Abuse Treatment Services. BHSIS Series S-84, HHS Publication No. (SMA) 16-4986. Rockville, MD: Substance Abuse and Mental Health Services Administration. Available at http://www.samhsa.gov/data/sites/default/files/2014_Treatment_Episode_Data_Set_National_Admissions_9_19_16.pdf

Volkow ND, Frieden TR, Hyde PS, Cha SS. Medication-assisted therapies--tackling the opioid-overdose epidemic. N Engl J Med. 2014 May 29;370(22):2063-6

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 (4b) under Usability and Use.*

The measure scores are calculated based on commercial claims data for two-year rolling periods for 2010-2015. The overall measure score results are shown in Table 1 and the results by state and health plan are shown in Tables 2 and 3, respectively. The number of episodes in the denominator of the measure ranged from a low of 17,229 in 2010-2011 to a high of 43,812 in 2013-2014 (Table 1). Over the period from 2010-2015, measure scores increased from 0.245 to 0.305.

Over the 2010-2015 time period, the number of states with at least 20 individuals eligible for the denominator increased from 44 states in 2010-2011 to 47 states in 2012-2013 (Table 2). Over the 2010-2015 time period, the number of health plans with at least 20 individuals eligible for the denominator increased from 88 health plans in 2010-2011 to 290 health plans in 2013-2014 (Table 3).

Table 1. Denominator, Numerator, and Measure Score for Two-Year Rolling Periods, 2010-2015

Time Period	Denominator	Numerator	Score
2010-2011	17,229	4,225	0.245
2011-2012	34,879	8,121	0.233
2012-2013	41,867	11,462	0.274
2013-2014	43,812	12,380	0.283
2014-2015	40,379	12,290	0.305

Table 2. Summary Statistics for Measure Scores by State, Two-Year Rolling Periods, 2010-2015

Time Period	Number of States	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
2010-2011	44	0.250	0.231	0.034	0.542	0.086	0.112	0.169	0.188	0.231	0.300	0.333
2011-2012	47	0.246	0.242	0.115	0.378	0.066	0.105	0.169	0.189	0.242	0.294	0.342
2012-2013	47	0.286	0.283	0.152	0.434	0.072	0.116	0.189	0.229	0.283	0.345	0.387
2013-2014	46	0.287	0.280	0.149	0.455	0.076	0.114	0.199	0.235	0.280	0.349	0.394
2014-2015	46	0.307	0.308	0.195	0.505	0.071	0.093	0.218	0.256	0.308	0.348	0.395

Table 3. Summary Statistics for Measure Scores by Health Plan, Two-Year Rolling Periods, 2010-2015

Time Period	Number of Health Plans	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
2010-2011	88	0.225	0.198	0.034	0.571	0.109	0.140	0.100	0.152	0.198	0.292	0.396
2011-2012	201	0.208	0.200	0.033	0.500	0.088	0.112	0.100	0.145	0.200	0.258	0.333
2012-2013	279	0.245	0.238	0.000	0.550	0.105	0.143	0.122	0.167	0.238	0.310	0.382
2013-2014	290	0.254	0.241	0.045	0.600	0.095	0.129	0.138	0.187	0.241	0.316	0.381
2014-2015	264	0.277	0.263	0.025	0.652	0.103	0.137	0.162	0.202	0.263	0.339	0.409

STD=standard deviation; IQR=interquartile range; PNN=NNth percentile

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.**
Data from the testing of the measure as specified are provided in 1b.2 above.

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 (4b) under Usability and Use.*
The measure was stratified for disparities by age and gender. The results/scores for 2013-2014 are presented for these categories in Table 4. The measure scores were lowest in individuals 18-34 years of age and highest in individuals 35-44 years of age. Scores were higher for males than females.

Table 4. Measure Scores by Age and Gender for Entire Sample, 2013-2014

Category	/ Denominator	/ Numerator	/ Measure Score
Age			
18-64 years	/ 43,812	/ 12,380	/ 0.283
18 –34	/ 25,360	/ 6,481	/ 0.256
35 –44	/ 9,008	/ 3,097	/ 0.344
45 –54	/ 6,294	/ 1,914	/ 0.304
55 – 64	/ 3,150	/ 888	/ 0.282
Gender			
Both Genders	/ 43,812	/ 12,380	/ 0.283
Female	/ 15,788	/ 4,229	/ 0.268
Male	/ 28,024	/ 8,151	/ 0.291

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**
Data on disparities from the testing of the measure as specified are provided in 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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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.)*

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_3175_OUD_Code_Lists_1-12-17_To_NQF.xlsx

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

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

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
Individuals in the denominator who have at least 180 days of continuous pharmacotherapy with a medication prescribed for OUD without a gap of more than seven days

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 measure numerator is calculated based on commercial claims data for rolling two-year periods from 2010 to 2015: 2010-2011, 2011-2012, 2012-2013, 2013-2014, and 2014-2015. The measure numerator is defined as individuals in the denominator with at least 180 days of "continuous pharmacotherapy" with an OUD medication.

Continuous pharmacotherapy for OUD is identified on the basis of the days covered by the days' supply of all prescription claims for any OUD medication (see list below) or number of days for which the drug was dispensed in a physician office or treatment center with the exceptions noted in this paragraph. The period of continuous pharmacotherapy starts on the day the first claim for an OUD medication is filled/supplied (index date) and lasts through the days' supply of the last claim for an OUD medication. To meet the 180-day requirement and be eligible for the measure, the date on the first claim for an OUD medication must fall at least 180 days before the end of the measurement period. For 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 two or more prescription claims occur on the same day or overlap, the surplus based on the days' supplies accumulates over all prescriptions. However, if another claim is submitted after a claim for an injectable OUD medication or an oral OUD medication

that is dispensed in an office or treatment center, the surplus from the day's supply for the injectable or office-dispensed medication is not retained.

An individual is considered to have continuous pharmacotherapy with OUD medication if there is no treatment gap of more than seven days. A gap is defined as a period during which the individual does not have oral OUD medication available based on the days' supply, or is more than 7 days overdue for having an injection of an extended-release OUD medication.

OUD medications were identified using National Drug Codes (NDCs) for the following:

- Buprenorphine
- Naltrexone (oral)
- Buprenorphine and Naloxone

And HCPCS codes for the following:

- Buprenorphine or Buprenorphine/naloxone, oral
- Methadone administration
- Naltrexone (extended-release injectable)

The National Drug Codes (NDCs) for the oral medications and the HCPCS codes for the injectable medications and office-dispensed oral medications (methadone and buprenorphine/naloxone) are contained in the sheets called "NDCs" and "HCPCS Codes", respectively, in the Excel file called "NQF 3175 OUD Code Lists" which is attached to this form under Item S.2b. Note that the NDC code list DOES NOT include NDC codes for methadone, as it can legally only be dispensed as OUD pharmacotherapy in licensed treatment centers. Buprenorphine can be dispensed through a pharmacy or in an office and is therefore identified based on either NDC or HCPCS codes.

Justification of Measure Definition: We define treatment continuity as (1) receiving at least 180 days of treatment and (2) no gaps in medication use of more than 7 days.

Our definition of minimum duration is based on the fact that the FDA registration trials for OUD drugs studied the effect of treatment over three to six months (US FDAa, undated; US FDAb, undated), and we have no evidence for effectiveness of shorter durations. In addition, several recommendations support a minimum six-month treatment period as the risk of relapse is the highest in the first 6-12 months after start of opioid abstinence (US FDAa, undated; US FDAb, undated; US DHHS, 2015). Longer treatment duration is associated with better outcomes compared to shorter treatments and the best outcomes have been observed among patients in long-term methadone maintenance programs ("Effective medical treatment of opiate addiction", 1998; Gruber et al., 2008; Moos et al., 1999; NIDA, 1999; Ouimette et al., 1998; Peles et al., 2013). Studies with long-term follow-up suggest that ongoing pharmacotherapy is associated with improved odds of opioid abstinence (Hser et al., 2015; Weiss et al., 2015). We did not specify a maximum duration of treatment, as no upper limit for duration of treatment has been empirically established (US DHHS, 2015).

We opted for using a treatment gap of more than seven days in our definition, given that the measure includes three active ingredients with different pharmacological profiles. There is substantial evidence for an elevated mortality risk immediately after treatment cessation (Cornish et al., 2010; Cousins et al., 2016; Davoli et al, 2007; Degenhardt et al., 2009; Gibson & Degenhardt, 2007; Pierce et al., 2016). Research suggests that methadone tolerance is lost after three days and this three-day threshold has been used in other observational methadone studies and in developing a United Kingdom treatment guideline which recommends revaluating patients for intoxication and withdrawal after a three-day methadone treatment gap (Cousins et al., 2016; Cousins et al., 2011; "Drug Misuse and Dependence—Guidelines on Clinical Management", 1999). Across all the medications, the mortality risk is highest in the first four weeks out of treatment, with many studies showing an increase in mortality in days 1-14 after treatment cessation.

Citations

Cornish R, Macleod J, Strang J, Vickerman P, Hickman M. Risk of death during and after opiate substitution treatment in primary care: prospective observational study in UK General Practice Research Database. BMJ. 2010;341:c5475.

Cousins G, Teljeur C, Motterlini N, McCowan C, Dimitrov BD, Fahey T. Risk of drug-related mortality during periods of transition in methadone maintenance treatment: a cohort study. J Subst Abuse Treat 2011; 41: 252–60.

Cousins G, Boland F, Courtney B, Barry J, Lyons S, Fahey T. Risk of mortality on and off methadone substitution treatment in primary care: a national cohort study. Addiction. 2016;111(1):73-82.

Davoli M, Bargagli AM, Perucci CA, et al. Risk of fatal overdose during and after specialist drug treatment: the VEdeTTE study, a national multisite prospective cohort study. Addiction. 2007;102:1954-9.

Degenhardt L, Randall D, Hall W, Law M, Butler T, Burns L. Mortality among clients of a state-wide opioid pharmacotherapy program over 20 years: risk factors and lives saved. Drug and alcohol dependence. 2009;105:9-15.

"Drug Misuse and Dependence—Guidelines on Clinical Management." Scottish Office Department of Health, Welsh Office, Social Services Northern Ireland. London: Stationery Office, 1999.

Effective medical treatment of opiate addiction. National Consensus Development Panel on Effective Medical Treatment of Opiate Addiction. JAMA.1998;280:1936-1943.

Gibson AE, Degenhardt LJ. Mortality related to pharmacotherapies for opioid dependence: a comparative analysis of coronial records. Drug Alcohol Rev. 2007; 26(4), 405-410.

Gruber VA, Delucchi KL, Kielstein A, Batki SL. A randomized trial of 6-month methadone maintenance with standard or minimal counseling versus 21-day methadone detoxification. Drug and Alcohol Dependence. 2008;94(1-3):199-206.

Hser YI, Evans E, Grella C, Ling W, Anglin D. Long-term course of opioid addiction. Harvard Review of Psychiatry. 2015;23(2):76-89.

Moos RH, Finney JW, Ouimette PC, Suchinsky RT. A comparative evaluation of substance abuse treatment: I. Treatment orientation, amount of care, and 1-year outcomes. Alcohol Clin Exp Res. 1999;23(3):529-36.

National Institute on Drug Abuse (NIDA). Principles of Drug Addiction Treatment: A Research-Based Guide. NIH Publication No. 99–4180. Rockville, MD: NIDA, 1999, reprinted 2000

Ouimette PC, Moos RH, Finney JW. Influence of outpatient treatment and 12-step group involvement on one-year substance abuse treatment outcomes. J Stud Alcohol. 1998;59:513-522

Peles E, Schreiber S, Adelson M. Opiate-dependent patients on a waiting list for methadone maintenance treatment are at high risk for mortality until treatment entry. J Addict Med. 2013;7(3):177-82..

Pierce M, Bird SM, Hickman M, Marsden J, Dunn G, Jones A, et al. Impact of treatment for opioid dependence on fatal drug-related poisoning: a national cohort study in England. Addiction. 2016;111:298-308.

U.S. Department of Health and Human Services Assistant Secretary for Planning and Evaluation Office of Disability, Aging and Long-Term Care Policy. Review of Medication-Assisted Treatment Guidelines and Measures for Opioid and Alcohol Use. Washington, DC, 2015. Accessed November 9, 2016 at: https://aspe.hhs.gov/sites/default/files/pdf/205171/MATguidelines.pdf

U.S. Food and Drug Administration (FDA) (a). REVIA Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/018932s017lbl.pdf

U.S. Food and Drug Administration (FDA) (b). VIVITROL Label. Accessed November 24, 2016 at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2006/021897lbl.pdf

Weiss RD; Potter JS; Griffin ML, et al. Long-term outcomes from the National Drug Abuse Treatment Clinical Trials Network Prescription Opioid Addiction Treatment Study. Drug and Alcohol Dependence. 2015;150:112-119.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Individuals 18-64 years of age who had a diagnosis of OUD and at least one claim for an OUD 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).
The measure denominator is calculated for rolling two-year periods from 2010 to 2015: 2010-2011, 2011-2012, 2012-2013, 2013-2014, and 2014-2015. The denominator includes individuals 18-64 years of age during their treatment period who had a diagnosis code of OUD during an inpatient, intensive outpatient, partial hospitalization, outpatient, detoxification or emergency department encounter at any time during the measurement period. To meet the 180-day requirement and be eligible for the measure, the date on the first claim for an OUD medication must fall at least 180 days before the end of the measurement period.

The diagnosis codes used to identify individuals with OUD included:
- ICD-9: 304.0x, 305.5x
- ICD-10: F11.xxx

These codes and descriptions are contained in the sheets called "ICD-9 Diagnosis Codes" and "ICD-10 Diagnosis Codes" in the Excel file called "NQF 3175 OUD Code Lists" which is attached to this form under Item S.2b.

OUD medications were identified using National Drug Codes (NDCs) for the following:
- Buprenorphine
- Naltrexone (oral)
- Buprenorphine and Naloxone

And HCPCS codes for the following:
- Buprenorphine or Buprenorphine/naloxone, oral
- Methadone administration
- Naltrexone (extended-release injectable)

The National Drug Codes (NDCs) for the oral medications and the HCPCS codes for the injectable medications and office-or treatment-center dispensed oral medications (methadone and buprenorphine) are contained in the sheets called "NDCs" and "HCPCS Codes", respectively, in the Excel file called "NQF 3175 OUD Code Lists" which is attached to this form under Item S.2b. Note that the NDC code list DOES NOT include NDC codes for methadone, as it can legally only be dispensed as OUD pharmacotherapy in licensed treatment centers. Buprenorphine can be dispensed through a pharmacy or in an office/treatment center and is therefore identified based on either NDC or HCPCS codes.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
There are no denominator exclusions.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
There are no denominator exclusions.

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.)*
Measure results may be stratified by:
- Age – Divided into four categories: 18-34, 35-44, 45-54, 55-64 years
- Gender: Male, Female
- State
- Health plan

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
The measure score is calculated for rolling two-year periods from 2010 to 2015. The steps described below are repeated for five rolling two-year periods: 2010-2011, 2011-2012, 2012-2013, 2013-2014, and 2014-2015. We present detailed results in the MIF for 2013-2014, as we have the most data for this time period, but we include measure scores for each of the two-year periods within 2010-2015.

DENOMINATOR: Individuals 18-64 years of age who had a diagnosis of OUD and at least one claim for an OUD medication

CREATE DENOMINATOR:
1. For each two-year period, identify individuals who are 18-64 years of age for the duration of the first year during which they appear in the period.

2. Of individuals identified in Step 1, keep those who had at least one encounter with any diagnosis (primary or secondary) of OUD in an outpatient setting, acute inpatient setting, or emergency department setting at any time during the two-year measurement period. The OUD diagnosis codes with descriptions are contained in the sheets called “ICD-9 Diagnosis Codes” and “ICD-10 Diagnosis Codes” in the Excel file called “NQF 3175 OUD Code Lists”, which is attached to this form under Item S.2b.

3. Of individuals identified in Step 2, keep those who have at least one claim with a National Drug Code (NDC) for any of the following oral OUD medications during the two-year period with a date at least 180 days before the end of the final calendar year of the measurement period:
- Buprenorphine
- Naltrexone (oral)
- Buprenorphine and Naloxone

Or a HCPCS code for any of the following OUD medications:
- Buprenorphine or Buprenorphine/naloxone, oral
- Methadone administration
- Naltrexone (extended-release injectable)

Claims for oral medications with negative, missing, or zero days’ supply were not included. The NDCs for the oral medications and the HCPCS codes for the injectable and office- or treatment center-dispensed medications are contained in the sheets called “NDCs” and “HCPCS Codes”, respectively, in the Excel file called “NQF 3175 OUD Code Lists,” which is attached to this form under Item S.2b.

4. Of individuals identified in Step 3, keep individuals who were continuously enrolled in a commercial health plan captured by our data for at least 6 months after the month with the first OUD medication claim in the measurement period, with no gap in enrollment. Individuals who are not enrolled for 6 months, including those who die during the period, are not eligible and are not included in the analysis. This is the denominator.

NUMERATOR: Individuals in the denominator who have at least 180 days of continuous pharmacotherapy with a medication prescribed for OUD without a gap of more than seven days

CREATE NUMERATOR:
For the individuals in the denominator, identify those who have at least 180 days of continuous pharmacotherapy with an OUD medication without a gap of more than seven days using the following method:

1. Determine the number of days for the PDC denominator. The start date is the service date (fill date) of the first prescription or injection/dispensing claim for an OUD medication in the two-year measurement period. The end date is defined as the earliest of:

- The date on which the individual exhausts their days' supply, including any pre-existing surplus, following their final claim (assuming daily use).
- The individual's death date.
- December 31st of the second year in the two-year period.

2. For each individual: Count the days during the observation period for which the individual was covered by at least one OUD medication based on the prescription drug or injection/dispensing claim service dates and days' supply.

2a. Sort OUD medication claims by individual's ID and service date. Scan the claims in order, calculating a rolling surplus which accumulates any remaining days' supply from other prior or same-day fills.
2b. Naltrexone injections contribute 30 days' supply unless another claim is found sooner, in which case the Naltrexone injection covers only the days up to the next claim.
2c. Methadone and buprenorphine/naloxone supply is determined by the start and end dates on the outpatient claims with the codes for in-office/treatment center dispensation of methadone (H0020) and buprenorphine/naloxone (J0571-J0575).
2d. Claims for Naltrexone injections and for licensed treatment center-dispensed methadone and office-dispensed buprenorphine/naloxone are not added to the surplus supply and only one such claim per day is counted.
2e. For 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.

3. Determine treatment gaps as periods, in which the individual has exhausted his/her available supply, defined as the days' supply from the most recent previous fill/dispensing and any pre-existing surplus available before that fill/dispensing.

4. Of the individuals in Step 2, count the number of individuals who have a period of 180 days or greater from the start date of the first claim for OUD medication to the end date of the last claim for OUD medication within the two-year period and who do not have a gap of more than seven days without OUD medication available. This is the numerator.

CALCULATE MEASURE SCORE:

1. Calculate the measure score by dividing the numerator by the denominator.

2. Calculate the measure score for each state. The state code on the claim record is used to identify individuals in each state. The measure score is then reported for each state that has at least 20 individuals in the denominator.

3. Calculate the measure score for each health plan. Health plan membership is approximated based on a combination of two variables found on the claim record, industry type and Metropolitan Statistical Area (MSA). A health plan identifier is assigned based on each unique combination of industry and MSA. The health plan identifier is used to group individuals into health plans. The measure score is then reported for each health plan that has at least 20 individuals in the denominator.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable; 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.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable; this measure does not use a sample or survey.

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

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
For measure calculation, the following files from the Truven MarketScan® Commercial Database were used:

- Enrollment data
- Drug claims
- Medical claims

We used data from these files (including data from Standard Quarterly Updates) for calendar years 2010-2015. This database has long been a commonly used data source to study patterns of commercially insured patients. The database contains fully adjudicated, patient-level claims. All records in these files were used as input to identify individuals that met the measure's eligibility criteria. We present detailed results in the MIF for 2013-2014, as we have the most data for this time period, but we include measure scores for each of the two-year periods within 2010-2015. The final analytic file for 2013-2014 contained a total of 43,812 episodes.

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, Population : Regional and State

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health : Outpatient, Clinician Office/Clinic
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
NQF_3175_OUD_Testing_Form_1-12-2017_To_NQF.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. (Do not remove prior testing information – include date of new information in red.)

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. (Do not remove prior testing information – include date of new information in red.)

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing

attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Measure Number (*if previously endorsed*)**:** 3175
Measure Title: Continuity of Pharmacotherapy for Opioid Use Disorder
Date of Submission: 1/12/2017
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes

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

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

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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

For measure calculation, the following files from the Truven MarketScan® Commercial Database were used:

- Enrollment data
- Drug claims
- Medical claims

We used data from these files (including data from Standard Quarterly Updates) for calendar years 2010-2015 for two-year rolling measure scores. This database has long been a commonly used data source to study patterns of commercially insured patients. The database contains fully adjudicated, patient-level claims. All records in these files were used as input to identify individuals that met the measure's eligibility criteria. We present detailed results in the MIF and this testing form for 2013-2014, as we have the most data for this time period, but we include measure scores for each of the two-year periods within 2010-2015. We restricted the sample to members with continuous enrollment of at least 180 days. The final analytic file for 2013-2014 contained a total of 43,812 episodes.

1.3. What are the dates of the data used in testing? January 1, 2010 – December 31, 2015

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.26*)	Measure Tested at Level of:
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☒ other: 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*)

Characteristics of the 2013-2014 denominator are summarized in Table 1 for the two levels of analysis, state and health plan. The sample for the state analysis included 46 states with 20 or more members eligible for the denominator.

As the data do not contain an actual health plan identifier, we developed a method based on the fact that the claims data are sourced from self-insured employers. We approximated health plan membership based on a combination of variables for industry type and Metropolitan Statistical Area (MSA) and assigned identifiers based on each unique

combination of industry and MSA. The sample for the health plan analysis included 290 health plans with 20 or more members eligible for the denominator.

Table 1. Denominator Characteristics by States and Health Plans, 2013-2014

Characteristics	States (n=46)	Health Plans (n=290)
Mean number per unit	928	52
Median number per unit	613	38
Minimum number per unit	43	20
Maximum number per unit	4,372	574
Standard Deviation	908.9	47.6
P10	66	21
P25	317	25
P50	613	38
P75	1,415	61
P90	1,904	96

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*)

Demographic characteristics of the individuals in the 2013-2014 dataset are shown in Table 2. For both the state and health plan analyses, more than half of the episodes fall into the 18-34 year age group, and almost two-thirds were male.

Table 2. Number Included in Testing of OUD Measure for States and Health Plans, by Demographic Characteristics, 2013-2014

Characteristic	States (n=46)	Health Plans (n=290)
Total	42,697	14,960
Age	Percent	Percent
18-34	57.8	57.8
35-44	20.5	19.8
45-54	14.4	14.5
55-64	7.2	7.9
Gender		
Female	35.9	36.4
Male	64.1	63.6

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.

There were no differences in the data used for different aspects of testing (e.g., measure scores, reliability).

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

The patient-level sociodemographic (SDS) variables that were available and analyzed in the data were age and gender.

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*)

The method of reliability testing used and the rationale are described below.

Method of Reliability Testing and Rationale

In order to assess measure precision in the context of the observed variability across measurement units (states and health plans), 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 (2009): "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."

This approach has two basic assumptions:

3. Each measured entity has a true pass rate, p, which varies between units of measurement following an unknown distribution between 0 and 1; and,
4. A sample proportion, calculated from a binomial random sample conditional on the measured entity's true pass rate.

As defined by Adams (2009), signal is defined as the variance in true pass rate between units of measurement, noise is defined as the estimation variance for each measured entity's score, and reliability scores are a ratio of signal to the sum of signal and noise.

We used the robust Prasad-Rao estimator for estimating the signal, and the standard binomial distribution inference for estimating the noise. Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual unit variance); whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across units). 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 units of observation.

Calculations were based on the mean denominator size for states (n=928) and health plans (n=52). 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."

Only health plans and states with more than one patient in the denominator were included in the calculation since units with only one observation cannot show any within-unit variation.

CITATIONS

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*)

Reliability scores were calculated at the state level and health plan level for the2013-2014 data.

The reliability score and standard deviation at the state level are 0.977 and 0.008, respectively, for the 2013-2014 data. This reliability score is greater than 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between states.

The reliability score and standard deviation at the health plan level are 0.891 and 0.040, respectively, for the 2013-2014 data. This reliability score is greater than 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between health plans.

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 results indicated that the measure scores were reliable at the state and health plan level, with both sets of measure scores having a reliability score of greater than 0.7, which is considered an acceptable cutpoint for adequate reliability (Scholle et al., 2008).

2b2. VALIDITY TESTING

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

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

☐ **Performance measure score**

☐ **Empirical validity testing**

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

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

We identified ten clinical experts in the treatment of OUD to rate the measure on its face validity and usability using a web-based questionnaire (developed using SurveyMonkey®). The names and organizations of the clinical experts are listed in Table 3.

The clinical experts were asked to review the measure specifications and the evidence supporting the measure from the NQF forms, which we provided to them. After reviewing the background material, they were instructed to rate two statements about the measure by indicating their level of agreement on a 5-point scale (1=Strongly Disagree; 2=Disagree; 3=Neither Agree nor Disagree; 4=Agree; 5=Strongly Agree). The first statement was related to the face validity of the measure: "Performance scores resulting from the measure as defined can be used to distinguish good from poor quality." The second statement was related to the usability of the measure: "The measure results are easily understood by the users of the data (e.g., clinicians, administrators)."

Table 3. Clinical Experts Who Rated Measure on Face Validity and Usability

Name	Affiliations and Employment
Adam Bisaga, MD New York, NY	Professor of Psychiatry Columbia University College of Physicians & Surgeons
Mady Chalk, PhD, MSW Philadelphia, PA	Managing Director, The Chalk Group Senior Policy Advisor, Treatment Research Institute, Philadelphia, PA
Bowen Chung, MD, MSHS Santa Monica, CA	Attending Physician Department of Psychiatry, Harbor-UCLA Medical Center Psychiatrist County of Los Angeles Department of Mental Health Associate Professor-in-Residence, Department of Psychiatry and Bio-behavioral Sciences David Geffen School of Medicine at UCLA Adjunct Scientist RAND Corporation
Louisa Degenhardt, PhD Sydney, Australia	Professor of Epidemiology NHMRC Principal Research Fellow Fellow of the Academy of Social Sciences of Australia
Keith Heinzerling, MD, MPH Los Angeles, CA	Associate Professor in Residence UCLA Department of Family Medicine
Brian Hurley, MD, MBA, DFASAM Los Angeles, CA	Addiction Psychiatrist Treasurer, American Society of Addiction Medicine Los Angeles County Department of Mental Health - Robert Wood Johnson Foundation Clinical Scholar at the David Geffen School of Medicine of the University of California, Los Angeles
Richard Saitz, MD, MPH, FACP, DFASAM Boston, MA	Chair, Department of Community Health Sciences (CHS) Professor of Community Health Sciences & Medicine Boston University School of Public Health
Jeffrey Samet, MD, MA, MPH Boston, MA	Professor of Medicine & Community Health Sciences Boston University Schools of Medicine & Public Health John Noble MD Professor in General Internal Medicine & Professor of Public Health Chief, General Internal Medicine, Boston Medical Center
Andrew Saxon, MD Seattle, WA	Professor and Director, Addiction Psychiatry Residency Program Department of Psychiatry & Behavioral Sciences University of Washington Director, Center of Excellence in Substance Abuse Treatment and Education (CESATE) VA Puget Sound Health Care System
Constance Weisner, DrPH, MSW Oakland, CA	Professor, Department of Psychiatry University of California, San Francisco Research Scientist, Division of Research Kaiser Permanente Northern California

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

Ten experts completed the evaluation of the measure's face validity and usability. The results of the rating of face validity on a scale of 1 to 5 are presented in Table 4.

Table 4. Results of the Face Validity Evaluation

Rating	Number with Rating (%)
5 (Strongly Agree)	1 (10%)
4 (Agree)	7 (70%)
3 (Neither Agree nor Disagree)	2 (20%)
2 (Disagree)	0 (0%)
1 (Strongly Disagree)	0 (0%)

Of the experts who rated the measure for face validity, 80 percent (8/10) strongly agreed or agreed with this statement: "Performance scores resulting from the measure as defined can be used to distinguish good from poor quality". The mean rating for face validity was 3.9, and the median rating 4.

The results of the rating of usability on a scale of 1 to 5 are presented in Table 5.

Table 5. Results of the Usability Evaluation

Rating	Number with Rating (%)
5 (Strongly Agree)	4 (40%)
4 (Agree)	5 (50%)
3 (Neither Agree nor Disagree)	0 (0%)
2 (Disagree)	1 (10%)
1 (Strongly Disagree)	0 (0%)

Of the experts who rated the measure for usability, 90 percent (9/10) strongly agreed or agreed with this statement: "The measure results are easily understood by the users of the data (e.g., clinicians, administrators)." The mean rating for usability was 4.2, and the median rating 4.

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

In summary, 80 percent of the OUD experts who participated in the rating strongly agreed or agreed that the measure has face validity, and 90 percent strongly agreed or agreed that the measure exhibits usability. This indicates strong support for the validity and usability of the measure.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)
Not applicable

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
Not applicable

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
Not applicable

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

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

2b4.1.1 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

2b4.2. 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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
Not applicable

2b4.4a. What were the statistical results of the analyses used to select risk factors?
Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Not applicable

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

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

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable

2b4.9. Results of Risk Stratification Analysis:
Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
Not applicable

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not applicable

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

To identify statistically significant differences in performance, we conducted a comparison of means and percentiles at the state and health plan level. Confidence intervals (95% CI) were calculated around point estimates for each state and health plan and then compared to the overall mean of states and proxy health plans, respectively. If the confidence intervals did not overlap with the overall mean, the difference was considered statistically significant.

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

We analyzed the 2013-2014 measure scores by state and health plan. The results for measure scores by state and health plan are presented in Table 6, along with a discussion of the meaningful differences at each level.

Table 6. Measure Score Performance at the State and Health Plan Level, 2013-2014

Level	n	Mean	Median	Min	Max	STD	IQR	P10	P25	P50	P75	P90
State	46	0.287	0.280	0.149	0.455	0.076	0.114	0.199	0.235	0.280	0.349	0.394

Health Plan	290	0.254	0.241	0.045	0.600	0.095	0.129	0.138	0.187	0.241	0.316	0.381

Meaningful Differences at the State Level – 2013-2014
In 2013-2014, 14 of 46 states (30.4 percent) had scores statistically significantly lower than the state-level mean, and 15 of 46 states (32.6 percent) had scores significantly higher than the state-level mean. For states with at least 20 episodes,tate-level measure scores ranged from a minimum of 0.149 to a maximum of 0.455, indicating suboptimal performance across all 46 states.

Meaningful Differences at the Health Plan Level – 2013-2014
In 2013-2014, 49 of the 290 health plans (16.9 percent) were statistically significantly lower than the health plan-level mean, and 2.8 percent of health plans (N=8) were statistically significantly higher than the health plan-level mean. For those health plans with at least 20 episodes, the scores for low- (10th percentile) and high- (90th percentile) performing plans were 0.138 and 0.381, respectively, indicating suboptimal performance across all health plans and wide variation between low- and high-performing health plans.

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

The results indicate that overall measure performance is suboptimal with variation in performance across states and health plans. Statistically significant differences were identified at the state and health plan level when compared to the overall mean.

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

One possible threat to validity is missing days' supply, which is a required data element to calculate the measure. An empirical assessment of this was conducted by analyzing the number (%) of individuals in a measure denominator in 2010-2015 with one or more claims that had missing days' supply.

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

Of 86,947 individuals in at least one of the 2010-2015 OUD cohorts, just 687 (0.8%) had one or more drug claims with a negative, zero, or missing value for days' supply. This small number of cases indicates that missing data do not pose a threat to the validity of the measure.

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

Threats to validity from missing data, to the extent we were able to address with testing, were not identified. The findings from the exploratory analysis suggest that very little impact on measure rates would be expected from 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.**
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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
The measure is not in operational use. Testing demonstrated that the measure was feasible to specify and calculate using administrative claims data. The claims data needed to implement the measure are available, accessible, and timely. Issues affecting feasibility regarding missing data were not identified. The cost of data collection is negligible, since the administrative data (collected primarily for billing purposes) are used as the data source for this measure. No other feasibility/implementation issues were identified.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees, licensing, or other requirements to use any aspect of the measure as specified.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization) Not in use	

4a.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

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

Not applicable; the measure is being submitted for initial endorsement.

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

Because the measure is being submitted to NQF for initial endorsement, we have not yet submitted it for use in a specific federal, state or local program. However, this measure would be appropriate for use in a Centers for Medicare & Medicaid Services (CMS) reporting program for Medicaid patients, such as the 2016 Core Set of Behavioral Health Measures for Medicaid. This list of 13 behavioral health measures was identified by CMS for voluntary reporting by state Medicaid and Children´s Health Insurance Program (CHIP) agencies. We will explore the possibility of submitting this measure through the Measures under Consideration (MUC) process for the one of the CMS reporting programs. This would entail submitting information about the measure through

JIRA, which is the CMS software system for collecting information on candidate measures for the list of "Measures under Consideration" for the annual pre-rulemaking process.

Improvement

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

4b. **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.
Not applicable; information about progress on improvement is not required because this measure is being submitted for initial endorsement.

4c. Unintended Consequences

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

4c.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
The measure has not been implemented in any reporting programs, and unexpected positive or negative findings were not identified during testing.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
The measure has not been implemented in any reporting programs, and therefore, unexpected benefits from implementation have not been observed.

4d1.1. **Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.**
How many and which types of measured entities and/or others were included? If only a sample of measured entities were included, describe the full population and how the sample was selected.
Not applicable; this measure is being submitted for initial endorsement.

4d1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
Not applicable; this measure is being submitted for initial endorsement.

4d2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
Not applicable; this measure is being submitted for initial endorsement.

4d2.2. **Summarize the feedback obtained from those being measured.**
Not applicable; this measure is being submitted for initial endorsement.

4d2.3. **Summarize the feedback obtained from other users**
Not applicable; this measure is being submitted for initial endorsement.

4d.3. **Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**

Not applicable; this measure is being submitted for initial endorsement.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

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

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

1664 : SUB-3 Alcohol & Other Drug Use Disorder Treatment Provided or Offered at Discharge and SUB-3a Alcohol & Other Drug Use Disorder Treatment at Discharge

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?

No

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

The target population of the proposed measure is related to the two measures listed above (NQF 0004 and NQF 1664). Differences among the three measures, along with the rationale and impact, are discussed below in the text box for Item 5b.1. The text box for this item (5a.2) would not accommodate the length of our response.

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 that address both the same measure focus and the same target population as the proposed measure.

RESPONSE TO ITEM 5A.2

The information below is the response to Item 5a.2, describing the differences, rationale, and impact on interpretability and data collection burden for the two NQF-endorsed RELATED measures which were identified. (We have inserted it here because the text box under Item 5a.2 would not accept this volume of formatted text.)

The target population of the proposed measure is related to the two NQF-endorsed measures listed above (NQF 0004 and NQF 1664). The proposed measure focuses on continuity of pharmacotherapy for patients with OUD. NQF 0004 focuses on treatment initiation and engagement of patients with a new episode of OUD or other substance use disorders, including alcohol use disorder (AUD). NQF 1664 focuses on 0UD and other drug use disorders among hospital discharges. Differences among the three measures, along with the rationale and impact are discussed below.

Diagnoses Included in Denominator Definition

- Proposed measure: Diagnosis of OUD
- NQF 0004: Diagnosis of alcohol or other drug dependence
- NQF 1664: Diagnosis of AUD or another substance use disorder
- Rationale and impact of focusing on only OUD: There are different medications for treatment of OUD and AUD, and there are no FDA-approved medications for treatment of other substance use disorders. In addition, the conceptual issues related to continuity of pharmacotherapy differ between OUD and AUD, so developing separate measures for the two disorders is required. The impact of this is a more narrowly focused measure that provides information specific to individuals with OUD.

Age Range

- Proposed measure: Patients 18-64 years of age
- NQF 0004: Patients aged 13 years of age and older
- NQF 1664: Patients 18 years of age and older
- Rationale and impact of limiting to individuals 18-64 years of age: Medications for treatment of OUD have not been approved by the FDA for adolescent patients 13-17 years of age; therefore, the proposed measure is restricted to adults 18-64 years of age.

Data Source

- Proposed measure: Electronic claims data
- NQF 0004: Administrative claims, electronic clinical data
- NQF 1664: Electronic clinical data, paper medical records
- Rationale and impact of using electronic claims data: Electronic claims data are timely, accessible, and relatively inexpensive to use for analyses of a large number of patients. Using a single source of data expedites the calculation of the measure, and will provide feedback to providers sooner.

Inpatient vs. Outpatient

- Proposed measure: Inpatient and outpatient
- NQF 0004: Inpatient and outpatient
- NQF 1664: Inpatient discharges
- Rationale and impact of using inpatient and outpatient records to identify patients: A large majority of patients with OUD are not admitted to a hospital, so using inpatient and outpatient data leads to more complete identification of the population eligible for treatment.

Process of Care Included in Numerator Definition

- Proposed measure: Continuity of pharmacotherapy for OUD
- NQF 0004: Inpatient admission, outpatient visit, intensive outpatient encounter, or partial hospitalization for adults with a new episode of AUD, OUD, or other substance use disorders
- NQF 1664: Medication for treatment of alcohol or drug use disorder OR a referral for addictions treatment
- Rationale and impact of the process of care included in the numerator definition: Successful pharmacotherapy of OUD requires continuity over at least a 180-day period. Therefore, providing feedback to providers about continuity of OUD pharmacotherapy has the potential to improve continuity rates by increasing provider awareness, and motivating health plans and insurers to develop educational material and programs about pharmacotherapy for OUD for both providers and patients.

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):** RAND Corporation
Co.2 **Point of Contact:** Soeren, Mattke, mattke@rand.org, 617-338-2059-8622
Co.3 **Measure Developer if different from Measure Steward:** RAND Corporation
Co.4 **Point of Contact:** Soeren, Mattke, mattke@rand.org, 617-338-2059-8622

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 group of ten behavioral health experts was used to rate the face validity and usability of the measure. Their names and affiliations are provided in the Testing Form.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Some proprietary codes are contained in the measure specifications for convenience of the user. Use of these codes may require permission from the code owner or agreement to a license.

Ad.7 **Disclaimers:** This performance measure does not establish a standard of medical care and has not been tested for all potential applications.

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

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 3185
Measure Title: Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention
Measure Steward: PCPI Foundation
Brief Description of Measure: Percentage of patients aged 18 years and older who were screened for tobacco use one or more times within 24 months AND who received cessation intervention if identified as a tobacco user
Developer Rationale: This measure is intended to promote adult tobacco screening and tobacco cessation interventions for those who use tobacco products. There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.

Numerator Statement: Patients who were screened for tobacco use at least once within 24 months AND who received tobacco cessation intervention if identified as a tobacco user
Denominator Statement: All patients aged 18 years and older seen for at least two visits or at least one preventive visit during the measurement period
Denominator Exclusions: Documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason)

Measure Type: Process
Data Source: Electronic Health Record (Only)
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.

This measure is the new eMeasure version of measure #3225(0028). The information provided for Evidence is identical to that submitted for #3225. Measure #3225 will be discussed first – the ratings for evidence will automatically be assigned to this eMeasure without further discussion. Also, because the claims/registry version of the measure currently is used in federal quality programs, BONNIE testing may be used as a source of synthetic test data to support eMeasure testing requirements.

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 #3225:

- The developer states the process/outcome rationale as: *There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.*
- Clinical practice guidelines from the U.S. Public Health Service (PHS) and recommendations statements from the U.S. Preventive Services Task Force (USPSTF) recommend that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products.

Updates:

- The following updated USPSTF (2015) statements support the components of this measure:
  - The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and U.S. Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. (“***A” recommendation, “good” or “fair” quality of evidence***)
  - The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. (“***A” recommendation, “good” or “fair” quality of evidence***)
- The developer summarizes the Quality, Quantity, and Consistency of evidence to be high across measure components. The developer indicated the review examined the impact of behavioral and pharmacologic interventions on 3 different outcomes:
  - health outcomes including mortality and morbidity
  - tobacco cessation
  - adverse events associated with tobacco cessation interventions

Exception to evidence: N/A

Guidance from the Evidence Algorithm

Process measure is based on a systematic review (SR) of the evidence and evidence is graded (Box 3) → Summary of QQC of the evidence provided (Box 4) → USPSTF Grade A, with Quantity: high; Quality: high; Consistency: high (Box 5a) → High

The highest possible rating is HIGH.

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.

- Average performance rates, based on data reported to the Physician Quality Reporting System (PQRS):
  - 2011: 81.6%
  - 2012: 84.1%
  - 2013: 89.7%
  - 2014: 88.9% (21.7% of eligible professionals reported on this measure)
- Note that the performance rates over time may not reflect rates from EHRs only.
- The average PQRS EHR performance rate for 2015 was 76.38%. The first decile rate is 27.84% and the third decile rate is 76.5%. Compared to the rates achieved when submitting data to PQRS via claims and registry data, providers submitting EHR data are not performing as well.

- The developers report that a number of studies have documented low rates of tobacco use screening and cessation intervention during primary care and other office/outpatient visits, missing key opportunities for intervention.

Disparities

- The developer indicates the federal reporting programs in which the measure is utilized have not made disparities data available for analysis and reporting.
- According to published data, disparities exist for counseling (less counseling for Hispanics vs. whites, those who are younger vs. older, and those with worker's compensation or unknown insurance vs. others) and for cessation assistance (higher for those with Medicaid/SCHIP insurance vs. those with private insurance or Medicare and for those who live in a high-poverty area vs. a low-poverty area).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

1.a. Evidence to Support Measure Focus

Comments:

**E-measure of 3225 so evidence is high.

**There is strong evidence that simple tobacco screening and brief smoking cessation interventions (counseling from the physician or pharmacotherapy) is an effective way to assist smokers to quit. Evidence is strong that quitting smoking directly reduces rates of lung cancer, heart disease, and stroke.

**1a. Evidence to Support Measure Focus:
-If measuring a structure, process, or intermediate outcome: How does the evidence relate to the specific structure, process, or intermediate outcome being measured?
There is good evidence for the measure.

-Does it apply directly or is it tangential? How does the structure, process, or intermediate outcome relate to desired outcomes?
It applies directly. If providers intervene then there is a good chance of the patient taking the advice or cessation medication.

-For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission?
I am not

-If measuring a health outcome or PRO: is the relationship between the measured outcome/PRO and at least one healthcare action (structure, process, intervention, or service) identified AND supported by the stated rationale?
N/A

1.b. Performance Gap

Comments:

** More room for improvement in the EHR performance rate. Disparity data same as 3225

**The studies presented show a gap in performance rates for this e-measure. While performance is rising related to EHRs there is still room for improvement warranting this e-measure.

Significant disparities exist related to use of smoking cessation interventions in minority communities, by age, by lower SES status, and region (urban vs. rural).

**1b. Performance Gap:
-Was performance data on the measure provided?
Yes. For those EPs reporting there is good performance, however, only 21.7 % of eligible EPs actually report on the measure in PQRS.

-How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure?
Many more providers need to adopt the measure

Disparities:
-Was data on the measure by population subgroups provided?
No as that data was not provided to the developer.

-How does it demonstrate disparities in the care?
Other data separate from the Measure Worksheet identify subgroups of the population that receive more or less counseling etc. see pg. 24

1.c. Composite
Comments:
**1c. Composite Performance Measure - Quality Construct (if applicable):
-Are the following stated and logical: overall quality construct, component performance measures, and their relationships; rationale and distinctive and additive value; and aggregation and weighting rules?

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

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

Data source(s): EHR electronic clinical data, electronic health record (EHR). This is an eMeasure.
Specifications:

- This measure is specified for the clinician group/practice level of analysis in the following settings: behavioral health outpatient; clinician office/clinic; home health; occupational therapy evaluation, speech and hearing evaluation, ophthalmological services visit. A higher score indicates better quality.
- The measurement period is a 24-month period.
- For the numerator, “tobacco use” includes any type of tobacco and a “tobacco cessation intervention” includes brief counseling (3 minutes or less) and/or pharmacotherapy.
- The denominator includes patients ages 18 or older, who have had at least 2 visits or 1 preventive care visit during the measurement period.
- Patients can be excluded from the denominator based on “medical reasons” for not screening (e.g., limited life expectancy). This should be coded using a CPT Category II code with modifier 4004F-1P.
- A calculation algorithm is included.
- The measure is not risk adjusted.
- HQMF specifications for the eMeasure are included in the document set on SharePoint. See the eMeasure Technical Review below.

Questions for the Committee:

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

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 [BONNIE testing] 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.

2a2. Reliability Testing, Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Reliability of the computed measure score was measured as the ratio of signal to noise using a beta-binomial model. This is an appropriate method of assessing score-level reliability.
  - A signal-to-noise analysis quantifies the amount of variation in performance that is due to differences between providers (as opposed to differences due to measurement error). Results will vary based on the amount of variation between the providers and the number of patients treated by each provider. The beta-binomial method typically results in a reliability statistic that ranges from 0 to 1 for each provider. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.
- The testing sample includes 2015 data reported via the EHR option to the PQRS program. This sample included data from 42,902 physicians, of whom 39,291 had all the required data elements and met the minimum number of quality reporting events (n=10). Data for 91.6% of reporting physicians were included in the reliability analysis.
- The developers provided 2 reliability estimates: reliability at the *minimum* number of events (n=10) and at the *average* number of events.
- NOTE that for testing, clinicians with <10 reporting events in the measurement period were excluded from the analysis. However, the measure specifications do not limit the measure to those with at least 10 reporting

events. This means that the reliability estimates from the analysis likely are higher than would be found if all clinicians were included (as typically, reliability increases with sample size).

Results of reliability testing (based on minimum threshold for inclusion=10 events)

Data source	Average number of events	Number of eligible providers meeting threshold	Percent of providers who did not meet threshold	Reliability at minimum number of events	Reliability at average number of events
EHR	524.5	39,291	8.4%	0.81	0.99

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm
Submitted specifications are precise, unambiguous and complete (Box 1) → Empirical reliability analysis conducted with measure as specified, except that testing was limited to providers with at least 10 patients (Box 2) → Reliability testing was conducted with computed performance measure score (Box 4) → Method was appropriate to assess variability in performance at measured entity level (Box 5) → Level of certainty that measure is reliable (Box 6) → Moderate

The highest possible rating is HIGH.

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

2b. Validity

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer conducted a face validity assessment by a 10-member expert panel. Participants included members of the newly-convened PCPI Preventive Care Technical Expert Panel. ***NOTE: This is the same assessment reported for measure #3225.***
  - After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement: "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." Agreement was measured based on a scale from 1 to 5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5= Strongly Agree.

- BONNIE testing also was conducted to test the measure logic and value sets for the e-Measure. This testing used a synthetic dataset with 40 patients. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. BONNIE testing output and screenshots are included in the document set on SharePoint (see NQF3185_FeasbilityReport).

Validity testing results:

- Face validity
  - N = 10
  - Mean rating = 3.6
  - 6 respondents (60%) either agreed or strongly agreed that this measure can accurately distinguish good and poor quality
- BONNIE testing
  - The testing results from the Bonnie tool reached 100% coverage and confirmed there was a test case for each pathway of logic (negative and positive test cases).
  - The measure also had a 100% passing rate which confirmed that all the test cases performed as expected.
  - The developer reports the measure logic performs as expected in the BONNIE system.

Questions for the Committee:

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

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients can be excluded from the denominator based on "medical reasons" for not screening (e.g., limited life expectancy).
- Based on 2015 PQRS data reported via EHRs: Among the 39,291 physicians with at least 10 quality reporting events:
  - 92,068 exceptions were reported
  - Average number of exceptions per physician= 2.3
  - Overall exception rate=0.4%
- The developers note that some have indicated concerns with this kind of "exception" reporting, including potential for gaming by inappropriately excluding patients in order to improve performance results. They cite Doran, et al, 2008 and Kmetik et al., stating that "research has indicated that levels of exception reporting occur infrequently and are generally valid".

Questions for the Committee:

- *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)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

- Data used in the analysis include 2015 PQRS data reported through EHRs.

Data source	Number of physicians	Mean	Standard Deviation	25th percentile	Median	75th percentile
EHRs	39,291	0.76	0.27	0.71	0.87	0.94

Question for the Committee:

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

2b6. Comparability of data sources/methods:

Not Applicable.

2b7. Missing Data

- No information was provided.

Guidance from the Validity Algorithm

Measure specifications are consistent with evidence provided (Box 1) → Potential threats to validity assessed (Box 2) → Empirical validity testing conducted (Box 3) → Face validity was systematically assessed (Box 4) and empirical testing of data elements was conducted using BONNIE tool (Box 10) → Method appropriate for legacy eMeasures (Box 11) → Moderate

The highest possible rating is MODERATE.

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

Committee pre-evaluation comments

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

2a.1 & 2b.1 Specifications: Reliability Specifications

Comments:

**Data elements are clearly defined.

**Why the denominator requires 2 visits during measurement period? Measure is not risk adjusted. EHRs are still inconsistently implemented in behavioral health locations across the country due to funding difficulties, reducing consistent adoption and implementation of this measure.

**2a1. & 2b1. Specifications:
Reliability-Specifications –
-Which data elements, if any, are not clearly defined?
They are clear, however, why is this only patients 18 and older? I could not find the reason for that other than where the measure was developed seemed to be in adult settings..

-Which codes with descriptors, if any, are not provided?
Codes seem appropriate

-Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear?
Seems logical

-What concerns do you have about the likelihood that this measure can be consistently implemented?

Will providers really remember to code the "Exceptions as 4004F-1P"?? see pg. 26 HER logic may help.

2a.2 Reliability Testing

Comments:

**Test sample is adequate to generalize with high percentage of physician data used and good reliability at both minimum and average events.

**The measure was tested on a large sample of providers (39,291). Data for 91% of the reporting physicians was included in the analysis. One caveat is that the reported reliability statistics are likely biased due to the exclusion of those not reporting a minimum of 10 events, even though the measure does not specify a limit of 10 events. Overall reliability appears very high (.81-.99) for those included in the sample. Given the large number in the sample, it would appear that the measure could be implemented reliably.

**2a2. Reliability - Testing:
-Was reliability tested with an adequate scope (number of entities and patients) to generalize for widespread implementation and with an appropriate method?
Seems this did not address younger people. Might be do to the setting being ones that focus on adults.

Would like to see younger people included as well. AHRQ has a measure that starts at age 12.
https://www.ahrq.gov/sites/default/files/wysiwyg/policymakers/chipra/factsheets/chipra_1516-p003-ef.pdf

-Describe how the results either do or do not demonstrate sufficient reliability.
Seemed to have good reliability testing.

If a PRO-PM: Was testing conducted at both the data element and score levels?

If a composite: Was testing conducted at the score level?

2b.1 Validity Specifications

Comments:

**yes

**Validity specifications appear consistent with the evidence presented for the measure.

**2b.1 Validity – Specifications:
-In what ways, if any, are the specifications inconsistent with the evidence?
They are consistent

-If a PRO-PM: In what ways, if any, are the specifications inconsistent with what the target population values and finds meaningful?

2b.2 Validity Testing

Comments:

**Face validity same as 3225, Bonnie testing 100%.

**Developer used face validity and BONNIE testing to examine measure validity. It appears that the results suggest that the measure is fairly reliable when it comes to face validity, and higher for the BONNIE method. It would appear that the e-measure is a fairly valid way to measure smoking screening and cessation interventions.

**2b2. Validity - Testing:
-Testing:

-Was validity tested with an adequate scope (number of entities and patients) to generalize for widespread implementation and with an appropriate method?
Again – only adults

-Describe how the results either do or do not demonstrate sufficient validity so that conclusions about quality can be made?

Seems to be valid for adults

-Why do you agree (or not agree) that the score from this measure as specified is an indicator of quality?
Asking, counseling and providing medication interventions have shown to be effective in getting people to decrease smoking.

-If a PRO-PM: Was testing conducted at both the data element and score levels?

2b3-7. Threats to Validity (Exclusions, Risk Adjustment, Statistically Significant Differences, Multiple Data Sources, Missing Data)
No adolescents included. I would like to see this resubmitted including adolescents.

2b.3.-2b7. Testing (Related to Potential Threats to Validity)
Comments:
**Same issue as with 3225

**Rather low rate of exceptions per physician, which would not appear to put an undue burden on clinicians. No risk adjustment discussion or justification provided. There do appear to be meaningful differences in the EHRs (.71-94), suggesting a possible need for quality improvement in this area.

**2b3. Exclusions:
-Are the exclusions consistent with the evidence?
For medical reasons yes. For age I do not think so.

-Are any patients or patient groups inappropriately excluded from the measure?
Yes Adolescents
-
Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?
I am not clear that in ambulatory settings you would have that many patients where smoking is OK due to a medical condition.

2b4. Risk Adjustment:
-If outcome (intermediate, health, or PRO-based) or resource use performance measure:

-Is there a conceptual relationship between potential SDS variables and the measure focus?
Yes. We know certain groups with specific SDS e.g. SMI patients have a higher ragte of smoking that the general population. This measure, however, is not risk adjusted.

-How well do SDS variables that were available and analyzed align with the conceptual description provided?
Developer did not provide.

-Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)?

-Was the risk adjustment (case-mix adjustment) appropriately developed and tested?

-Do analyses indicate acceptable results?

-Is an appropriate risk-adjustment strategy included in the measure?

2b5. Meaningful Differences:

-How do analyses indicate this measure identifies meaningful differences about quality?
Measure can help to identify providers that do focus on and assist patients to quit smoking.

2b6. Comparability of performance scores:
-If multiple sets of specifications:
-Do analyses indicate they produce comparable results?
-If risk-adjustment approach includes SDS factors:

Did the developer compare performance scores with and without SDS factors in the risk-adjustment approach?
Did the results support the risk-adjustment approach?

Not risk adjusted
2b7. Missing data/no response:
-Does missing data constitute a threat to the validity of this measure?
I do not think so.

2d. Composite Performance Measure
Comments:
**2d. Composite Performance Measure - Composite Analysis (if applicable):
-Do analyses demonstrate the component measures fit the quality construct and add value?
-Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

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.

- A 2016 feasibility assessment was performed, in order to assess the extent to which the required data are readily available and can be accurately captured in a standardized way, without undue burden.
- Two entities participated in the Feasibility assessment for this measure:
  - A 619-bed multispecialty academic medical center, serving 33 counties, in the state of California
  - A 200-bed acute care facility in Chicago, IL
  - One of the entities used the Epic and the other used the VA/VistA system.
- The assessment provided shows that measure logic performed as expected in the BONNIE system. Two feasibility score cards were provided (one for each entity that participated in testing).
  - In one system, all data elements were currently feasible.
  - In the other system, only 17 of the 26 data elements were currently feasible, but all were judged to be feasibility within the next 3-5 years.
- In summarizing their feasibility assessment, the developer states:
  - *"The required data elements are captured in the Electronic Health Record. A few encounter types are not coded in some systems. Measure exceptions and measure exception examples are currently not captured in a standardized format in some systems, but are able to be captured via free text."*

Questions for the Committee:
- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

3. Feasibility
Comments:
**The data strategy is ready for use. It is concerning that not all data elements are able to be captured by one EHR system used, although developer says will be remedied has a 3-5 year time frame. I was unable to open the feasibility scorecards.

**EHRs do collect smoking cessation information, but are not likely consistent across applications. There are still behavioral health providers, particularly smaller entities, that have not fully adopted EHRs; however this will continue to change with time.

**3. Feasibility:
-Which of the required data elements are not routinely generated and used during care delivery?
-Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)?

-What are your concerns about how the data collection strategy can be put into operational use?

Will providers really remember to code the "Exceptions as 4004F-1P"?? see pg. 26

Criterion 4: Usability and Use

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- This measure is included in the Physician Quality Reporting System, and publicly reported on Physician Compare.
- It is also used in the Million Hearts program.
- Although not mentioned in the submission materials, NQF staff believe this measure also is being used in the following CMS programs:
  - Medicare Shared Savings Program (MSSP)
  - Physician Value-Based Payment Modifier (VBM) [which is being phased out by 12/31/18 and is replaced by MIPS]
  - Physician Feedback/Quality and Resource Use Reports (QRUR) [which is being phased out by 12/31/18 and is replaced by MIPS]

Improvement results

- Trends in performance based on EHR (eMeasure) data likely were not provided. However, performance results for 2011-2014 were provided, based on data from EHRs, claims, and registries.

Unexpected findings (positive or negative) during implementation: None reported.

Potential harms: None indicated.

Vetting of the measure

- The developer does not discuss any provision of the results and data to those being measured.
- The developer obtains feedback from implementers and those being assessed through a variety of mechanisms.
- In response to feedback from implementers or others, some modifications to the measure specifications and guidance were made prior to finalizing the measure.

Feedback:

- The developers have a process to accept and respond to implementer comments and questions. They report that most questions are for clarification regarding what does, or does not, meet the measure. However, they

also note that "*More recently, many implementers wanted to understand how the measure addresses electronic nicotine delivery systems (ENDS*)".

***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?*
- *How has the measure been vetted in real-world settings by those being measure or others?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Usability and Use:

Comments:

**Per 3225.

**This measure can be used to improve the quality of practice. This measure is already included in a number of areas, including the Physician Quality Reporting System, and is publically reported. The benefits to the patients far outweigh any unintended consequences of its use.

**4. Usability and Use:
-How is the measure being publicly reported?

-For maintenance measures – which accountability applications is the measure being used for?
MACRA/QPP; NYS Medicaid and other programs

-How can the performance results be used to further the goal of high-quality, efficient healthcare?
Could be helpful in identifying providers that need assistance in incorporating the measures into practice.

-Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.
I do not see any.

Has the measure been vetted in real-world settings by those being measured or others?
-If so, has data, results, and aid in interpretation been provided?
-Has feedback been solicited?
-Was feedback considered if/when changes were made to the measure?
Yes

I do not understand why measure # 3225 previously NQF# 0028 is being given a new NQF #. Providers are already using 0028. If we change the NQF number they will lose their history of measure performance. I am not seeing any good rationale for this and would suggest keeping the 0028 # and making the eMeasure 0028e.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0027: Medical Assistance With Smoking and Tobacco Use Cessation
- 1651: TOB-1 - Tobacco Use Screening
- 1654: TOB-2 - Tobacco Use Treatment Provided or Offered
- 1656: TOB-3- Tobacco Use Treatment Provided or Offered at Discharge
- 2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
- 2803 : Tobacco Use and Help with Quitting Among Adolescents

- 3225: Tobacco Use: Screening and Cessation Intervention

Harmonization

- Measure #0027 is a health plan measure that assess patient-reported advice and advice to quit smoking and other tobacco use, as well as discussion of discussion of cessation strategies and medications. Data for this measure are obtained from the Health Plan CAHPS survey.
- Measures #1651, #1654, and #1656 are hospital-level measures aimed at offering/providing screening, counseling, and cessation interventions
- Measure 2600 focuses on specific populations (SMI, AOD) at the health plan level
- Measure 2803 looks at screening and cessation interventions in adolescents at the clinician group/practice level.
- Measure #3225 is the claims/registry version of this measure. It appears to be harmonized with #3185 to the extent possible.
- These measures seem to be mostly harmonized in terms of their definitions, but potential for further harmonization on specific measures may be discussed at the in-person meeting.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because empirical validity testing for the measure score using the HQMF specifications on data from EHRs has not been conducted.

Pre-meeting public and member comments

-

NQF #: 0028 NQF Project: Behavioral Health

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:

This measure is intended to promote adult tobacco screening and tobacco cessation interventions for those who use tobacco products. There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

The measure focuses on routine tobacco screening for all adults and tobacco cessation interventions for those who use tobacco products. Tobacco use includes use of any type of tobacco.

Clinical practice guidelines from the U.S. Public Health Service (PHS) and recommendations statements from the U.S. Preventive Services Task Force (USPSTF) recommend that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products. The PHS guideline noted that the majority of clinician attention and research in the field has focused on the treatment and assessment of smoking. Nevertheless, they indicated that "[t]he interventions found to be effective in this Guideline have been shown to be effective in a variety of populations. In addition, many of the studies supporting these interventions comprised diverse samples of tobacco users. Therefore, interventions identified as effective in this Guideline are recommended for all individuals who use tobacco, except when medication use is contraindicated or with specific populations in which medication has not been shown to be effective (pregnant women, smokeless tobacco users, light smokers, and adolescents)."

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

In 2015, the USPSTF published an update to its 2009 recommendation on counseling and interventions to prevent tobacco use and tobacco-related disease in adults, including pregnant women. Because there were no plans to update the Public Health Service clinical practice guidelines on treating tobacco use and dependence which formed the basis for the original USPSTF recommendation (2003) and reaffirmation (2009), the Agency for Healthcare Research and Quality (AHRQ) commissioned a new evidence review to assess the benefits and harms of behavioral and pharmacologic interventions for tobacco cessation in adults, including pregnant women. As a result, the 2015 USPSTF updated recommendation is based on the evaluation of evidence summarized in the 2015 review of reviews.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles*): Since the measure essentially addresses three components (ie, (1) screening and cessation interventions comprising (2) brief counseling and/or (3) pharmacotherapy), the quantity of studies noted by the guideline are offered as they relate to each of the measure components.

For screening and assessment and its impact on clinical intervention, 9 studies met the selection criteria and were meta-analyzed.

For screening and assessment and its impact on tobacco cessation, 3 studies met the selection criteria and were meta analyzed.

For advice to quit smoking, 7 studies were included in the meta-analysis. For specific information about the intensity of the intervention, namely the efficacy of minimal counseling interventions lasting less than 3 minutes in comparison to low-intensity or high-intensity counseling interventions, 43 studies met the selection criteria for comparison across various lengths.

For combining counseling and medication, 18 studies met selection criteria.

For medication alone, a meta-analysis of 83 studies evaluated the effectiveness and abstinence rates for various medications and medication combinations compared to placebo at 6-months post-quit.

2015 Review of Reviews for the USPSTF

As described above, the evidence review published in 2015 focused on the benefits and harms of behavioral and pharmacologic interventions for tobacco cessation in adults, including pregnant women. It relied primarily on a review of reviews method and included relevant reviews from January 2009 through August 1, 2014.

For behavioral interventions among adults, 26 systematic reviews were included in the analysis.

For pharmacotherapy interventions among adults, 9 systematic reviews were included in the analysis.

For combined pharmacotherapy and behavioral interventions among adults, 1 systematic review was included in the analysis.

For behavioral interventions among pregnant women, 6 systematic reviews were included in the analysis.

For pharmacotherapy interventions among pregnant women, 6 systematic reviews were included in the analysis.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of the body of evidence supporting each of the PHS guideline recommendations is summarized according to the strength of evidence ratings as "A." "A" evidence is described as "Multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings."

Additionally, the medication meta-analysis included predominantly studies with “self-selected” populations. In addition, in medication studies both experimental and control subjects in the studies typically received substantial counseling. Both of these factors tend to produce higher abstinence rates than typically are observed among self-quitters.

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

2015 Review of Reviews for the USPSTF

The quality of the evidence was rated by 2 independent reviewers using a slightly modified version of the AMSTAR (Assessment of Multiple Systematic Reviews) tool. The reviewers then applied the typical USPSTF quality scores (i.e., good-quality, fair-quality, or poor-quality) as described below:

- Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes.
- Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes.
- Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.

All poor quality studies were excluded from the analysis.

For behavioral interventions among adults, 16 systematic reviews were rated as good quality, 10 were rated as fair quality.

For pharmacotherapy interventions among adults, 5 systematic reviews were rated as good quality, 4 were rated as fair quality.

For combined pharmacotherapy and behavioral interventions among adults, 1 systematic review was rated as good quality.

For behavioral interventions among pregnant women, 3 systematic reviews were rated as good quality, 3 were rated as fair quality.

For pharmacotherapy interventions among pregnant women, 5 systematic reviews were rated as good quality, 1 was rated as fair quality.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The consistency of results across studies is summarized according to the strength of evidence ratings as "A." "A" evidence is described as "Multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings."

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

The magnitude and direction of the effect across studies is summarized below for each relevant component addressed by the PHS guideline.

2015 Review of Reviews for the USPSTF

In general, results across all included reviews were consistent within each population and intervention grouping. Reviews rated as good, by definition, include "consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes." The magnitude and direction of the effect for each population and intervention grouping is summarized below.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

For screening and assessment, the PHS panel looked at two different outcomes - the impact on clinical intervention and tobacco cessation. They concluded that "having a clinic system in place that identifies smokers increases rates of clinician intervention but does not, by itself, produce significantly higher rates of smoking cessation."

Results of the meta-analysis for advice to quit smoking show that brief physician advice significantly increases long-term smoking abstinence rates.

Results of the meta-analysis regarding the intensity of the counseling intervention revealed that all three session lengths (minimal counseling, low-intensity counseling, and higher intensity counseling) significantly increased abstinence rates over those produced by no-contact conditions.

However, there was a clear trend for abstinence rates to increase across these session lengths, with higher intensity counseling producing the highest rates.

For combining counseling and medication, the results of the meta-analysis indicate that providing counseling in addition to medication significantly enhances treatment outcomes.

For medication alone, the PHS Panel identified seven first-line (FDA-approved) medications (bupropion SR, nicotine gum, nicotine inhaler, nicotine lozenge, nicotine nasal spray, nicotine patch, and varenicline) and two second-line (non-FDA-approved for tobacco use treatment) medications (clonidine and nortriptyline) as being effective for treating smokers. Each has been documented to

increase significantly rates of long-term smoking abstinence. These medications should be encouraged except where contraindicated or for specific populations for which there is insufficient evidence of effectiveness (i.e., pregnant women, smokeless tobacco users, light smokers, and adolescents).

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

2015 Review of Reviews for the USPSTF

Where possible, the review examined the impact of behavioral and pharmacologic interventions on 3 different outcomes:

- health outcomes including mortality and morbidity
- tobacco cessation
- adverse events associated with tobacco cessation interventions

For behavioral interventions among adults:

- *Health Outcome:* 1 trial found favorable effects on all-cause and coronary disease mortality and lung cancer incidence and mortality 20 years after an intensive behavioral intervention, although results were not statistically significant.
- *Cessation Outcome:* Health provider advice and counseling, tailored self-help materials, and telephone counseling showed modest but significant increased smoking cessation at ≥6 months relative to control participants (18%–96%). Providing more intense adjunctive behavioral support to smokers receiving pharmacotherapy may increase cessation by 9%–24%. Evidence on the use of mobile phone support, Internet-based interventions, and complementary and alternative therapies was limited and not definitive
- *Adverse Event (AE):* Minor AEs related to ear acupuncture, ear acupressure, and other auriculotherapy have been reported. AEs related to other behavioral or complementary and alternative therapies have not been documented.

For pharmacotherapy interventions among adults:

- *Cessation Outcome:* NRT, bupropion SR, and varenicline improve the chances of smoking cessation. Reviews suggested that NRT might increase smoking abstinence at ≥6 mo by 53%–68%, bupropion SR by 49%–76%, and varenicline by 102%–155%. Absolute cessation differences averaged 7% for NRT, 8.2% for bupropion SR, and 26% for varenicline. There were no significant differences among different NRT products, and relative rates of abstinence were similar across settings. Use of a combination of NRT products increases cessation rates more than the use of a single NRT product. In general, there were no significant differences among different classes of medications in direct comparisons.
- *Adverse Event:* NRT, bupropion SR, and varenicline are not associated with an increased risk for major CV AEs. NRT is associated with a higher rate of any CV AE largely driven by low-risk events, typically tachycardia. There was a marginal, nonsignificant increase in serious AEs in participants receiving bupropion SR but no difference for serious psychiatric AEs. The evidence for the safety of varenicline is still under investigation; 1 review suggested a 36% increased risk for nonfatal serious AEs among those receiving varenicline vs. a control intervention.

For combined pharmacotherapy and behavioral interventions among adults,

- *Cessation Outcome:* Combined pharmacotherapy and behavioral interventions increase cessation rates by 70%–100% compared with no or minimal treatment.

For behavioral interventions among pregnant women:

- *Health Outcome:* Statistically significant benefit of behavioral interventions on mean birthweight, low birthweight, and preterm birth vs. usual care or control.

- *Cessation Outcome:* Pooled estimates of a range of behavioral interventions from 70 studies suggested benefits for validated smoking cessation, with a similar benefit when limited to the most common intervention (counseling). Heterogeneity was moderate for the pooled effect, but there was no evidence of subgroup effects by intervention type, number of intervention components, or outcome ascertainment approach.
- *Adverse Event:* No serious AEs reported.

For pharmacotherapy interventions among pregnant women

- *Health Outcome:* Limited evidence of NRT on perinatal and child health benefits. 3 of 4 NRT trials reported fewer preterm births in the intervention group, but only 1 was statistically less than placebo. 2 trials reported higher birthweight in the NRT group; 2 larger trials found no difference. Follow-up data from the largest NRT trial found a higher rate of "survival with no impairment" at 2 y among children of women assigned to the NRT intervention vs. placebo (73% vs. 65%). No trials of bupropion SR or varenicline among pregnant women.
- *Cessation Outcome* No statistical evidence of NRT efficacy for validated smoking cessation in late pregnancy, but power was limited and all trials were in the direction of benefit (pooled analysis based on 5 placebo-controlled trials). No trials of bupropion SR or varenicline among pregnant women.
- *Adverse Event:* No evidence of perinatal harms from NRT. 1 trial found a higher rate of cesarean section for women assigned to NRT; follow-up from the same trial was reassuring for child health outcomes. No trials of bupropion SR or varenicline among pregnant women.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

2015 Review of Reviews for the USPSTF

Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The PHS guideline is the product of a private-sector panel of experts

("the Panel"), representatives of a consortium of several Federal Government and nonprofit organizations, and staff. The panel membership included: Michael C. Fiore, MD, MPH (Panel Chair); Carlos Roberto Jaén, MD, PhD, FAAFP (Panel Vice Chair); Timothy B. Baker, PhD (Senior Scientist); William C. Bailey, MD, FACP, FCCP; Neal L. Benowitz, MD; Susan J. Curry, PhD; Sally Faith Dorfman, MD, MSHSA; Erika S. Froelicher, PhD, RN, MA, MPH;

Michael G. Goldstein, MD; Cheryl G. Healton, DrPH; Patricia Nez Henderson, MD, MPH; Richard B. Heyman, MD; Howard K. Koh, MD, MPH, FACP; Thomas E. Kottke, MD, MSPH; Harry A. Lando, PhD; Robert E. Mecklenburg, DDS, MPH; Robin J. Mermelstein, PhD; Patricia Dolan Mullen, DrPH; C. Tracy Orleans, PhD; Lawrence Robinson, MD, MPH; Maxine L. Stitzer, PhD; Anthony C. Tommasello, PhD, MS; Louise Villejo, MPH, CHES; Mary Ellen Wewers, PhD, MPH, RN.

The evaluation of conflict for the 2008 Guideline Update comprised a two-stage procedure designed to obtain increasingly detailed and informative data on potential conflicts over the course of the Guideline development process. Of the Panel members listed in this document, 21 of 24 had no significant financial interests as defined by the PHS-based criteria. In addition to these mandatory disclosures regarding compensation, leadership, and ownership, members were asked to disclose any other information that might be disclosed in a professional publication. Three Panel members whose disclosures exceeded the PHS criteria for significant financial interest were recused from Panel deliberations relating to their areas of conflict; one additional Panel member voluntarily recused himself.

2015 Review of Reviews for the USPSTF

At least 2 independent reviewers rated the quality of all included systematic review. Discrepancies were resolved through discussion. Additional information regarding disclosures is included in Section 1C.20 below

1c.11 System Used for Grading the Body of Evidence: Other

2015 Review of Reviews for the USPSTF: USPSTF (described in 1c.6.)

1c.12 If other, identify and describe the grading scale with definitions: Every recommendation made by the PHS Panel bears a strength-of-evidence rating that indicates the quality and quantity of empirical support for the recommendation. Each recommendation and its strength of evidence reflects consensus of the Guideline Panel.

The three strength-of-evidence ratings are described as follows:

A. Multiple well-designed randomized clinical trials, directly relevant to

the recommendation, yielded a consistent pattern of findings.

B. Some evidence from randomized clinical trials supported the recommendation, but the scientific support was not optimal. For instance, few randomized trials existed, the trials that did exist were somewhat inconsistent, or the trials were not directly relevant to the recommendation.

C. Reserved for important clinical situations in which the Panel achieved consensus on the recommendation in the absence of relevant randomized controlled trials.

1c.13 Grade Assigned to the Body of Evidence: PHS: A; USPSTF does not separately grade the body of evidence

1c.14 Summary of Controversy/Contradictory Evidence: No controversy or contradictory evidence reported.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:

PHS Guideline (1):

All patients should be asked if they use tobacco and should have their tobacco use status documented on a regular basis. Evidence has shown that clinic screening systems, such as expanding the vital signs to include tobacco use status or the use of other reminder systems such as chart stickers or computer prompts, significantly increase rates of clinician intervention. (Strength of Evidence = A)

All physicians should strongly advise every patient who smokes to quit because evidence shows that physician advice to quit smoking increases abstinence rates. (Strength of Evidence = A)

Minimal interventions lasting less than 3 minutes increase overall tobacco abstinence rates. Every tobacco user should be offered at least a minimal intervention, whether or not he or she is referred to an intensive intervention. (Strength of Evidence = A)

The combination of counseling and medication is more effective for smoking cessation than either medication or counseling alone. Therefore, whenever feasible and appropriate, both counseling and medication should be provided to patients trying to quit smoking. (Strength of Evidence = A)

Clinicians should encourage all patients attempting to quit to use effective medications for tobacco dependence treatment, except where contraindicated or for specific populations for which there is insufficient evidence of effectiveness (i.e., pregnant women, smokeless tobacco users, light smokers, and adolescents). (Strength of Evidence = A)

USPSTF Recommendation (2):

The USPSTF recommends that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products. This is a grade A recommendation.

USPSTF Recommendation (3):

The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and U.S. Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. (A recommendation)

The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. (A recommendation)

1c.17 Clinical Practice Guideline Citation: 1. Fiore MC, Jaen CR, Baker TB, et al. Treating tobacco use and dependence: 2008 update. Clinical practice guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.

2. U.S. Preventive Services Task Force. Counseling and interventions to prevent tobacco use and tobacco-caused disease in adults and pregnant women: U.S. Preventive Services Task Force reaffirmation recommendation statement. Ann Intern Med 2009 Apr 21;150(8):551-5.

3. Siu AL; U.S. Preventive Services Task Force. Behavioral and Pharmacotherapy Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Women: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2015 Oct 20;163(8):622-34. doi: 10.7326/M15-2023. Epub 2015 Sep 22.

1c.18 National Guideline Clearinghouse or other URL: www.surgeongeneral.gov/tobacco; www.uspreventiveservicestaskforce.org/uspstf/uspstbac2.htm

https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/tobacco-use-in-adults-and-pregnant-women-counseling-and-interventions1

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The Members of the U.S. Preventive Services Task Force members at the time the 2009 recommendation was finalized represented an array of health-related disciplines including internal medicine, family medicine, behavioral medicine, pediatrics, obstetrics/gynecology and nursing. The task force membership comprised the following individuals: Ned Calonge, MD, MPH, Chair (Colorado Department of Public Health and Environment, Denver, Colorado); Diana B. Petitti, MD, MPH, Vice-Chair (Arizona State University, Phoenix, Arizona); Thomas G. DeWitt, MD (Children´s Hospital Medical Center, Cincinnati, Ohio); Allen J. Dietrich, MD (Dartmouth Medical School, Hanover, New Hampshire); Kimberly D. Gregory, MD, MPH (Cedars-Sinai Medical Center, Los Angeles, California); David Grossman, MD (Group Health Cooperative, Seattle, Washington); George Isham, MD, MS (HealthPartners Inc., Minneapolis, Minnesota); Michael L. LeFevre, MD, MSPH (University of Missouri School of Medicine, Columbia, Missouri); Rosanne M. Leipzig, MD, PhD (Mount Sinai School of Medicine, New York, New York): Lucy N. Marion, PhD, RN (School of Nursing, Medical College of Georgia, Augusta, Georgia); Bernadette Melnyk, PhD, RN (Arizona State University College of Nursing & Healthcare Innovation, Phoenix, Arizona); Virginia A. Moyer, MD, MPH (Baylor College of Medicine, Houston, Texas); Judith K. Ockene, PhD (University of Massachusetts Medical School, Worcester, Massachusetts); George F. Sawaya, MD (University of California, San Francisco, San Francisco, California); J. Sanford Schwartz, MD (University of Pennsylvania Medical School and the Wharton School, Philadelphia, Pennsylvania); and Timothy Wilt, MD, MPH (University of Minnesota Department of Medicine and Minneapolis Veteran Affairs Medical Center, Minneapolis, Minnesota). Prior to each meeting, Task Force members are asked to disclose any information that may interfere with their abilities to discuss and/or vote on a specific topic. Conflicts may arise, for example, if a member has a financial, business/professional, and/or intellectual interest in areas related to a particular topic. All members are expected to provide full disclosure of their interests related to all topics that will be discussed at each meeting. A committee comprised of AHRQ staff and the USPSTF Chair and Vice Chair review each member´s disclosures and issue a recommendation on the member´s eligibility to participate on a specific topic(s). Each member is notified by AHRQ staff of the recommendation prior to each meeting. Members are free to recuse themselves voluntarily from participation in the processes for specific topics; however, a voluntary recusal does not free a member from the obligation to disclose a conflict.

USPSTF 2015 Recommendation Statement

Members of the USPSTF at the time this recommendation was finalized are Albert L. Siu, MD, MSPH, *Chair* (Mount Sinai School of Medicine, New York, and James J. Peters Veterans Affairs Medical Center, Bronx, New York); Kirsten Bibbins-Domingo, PhD, MD, MAS, *Co-Vice Chair* (University of California, San Francisco, San Francisco, California); David Grossman, MD, MPH, *Co-Vice Chair* (Group Health, Seattle, Washington); Linda Ciofu Baumann, PhD, RN, APRN (University of Wisconsin, Madison, Wisconsin); Karina W. Davidson, PhD, MASc (Columbia University, New York, New York); Mark Ebell, MD, MS (University of Georgia, Athens, Georgia); Francisco A.R. García, MD, MPH (Pima County Department of Health, Tucson, Arizona); Matthew Gillman, MD, SM (Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts); Jessica Herzstein, MD, MPH (Independent Consultant, Washington, DC); Alex R. Kemper, MD, MPH, MS (Duke University, Durham, North Carolina); Alex H. Krist, MD, MPH (Fairfax Family Practice, Fairfax, and Virginia Commonwealth University, Richmond, Virginia); Ann E. Kurth, PhD, RN, MSN, MPH (New York University, New York, New York); Douglas K. Owens, MD, MS (Veterans Affairs Palo Alto Health Care System, Palo Alto, and Stanford University, Stanford, California); William R. Phillips, MD, MPH (University of Washington, Seattle, Washington); Maureen G. Phipps, MD, MPH (Brown University, Providence, Rhode Island); and Michael P. Pignone, MD, MPH (University of North Carolina, Chapel Hill, North Carolina). Former USPSTF member Susan Curry, PhD, also contributed to the development of this recommendation.

The USPSTF requires each member to disclose all information regarding any possible financial and nonfinancial conflicts of interest prior to each meeting for all topics under development or that will be discussed at each meeting. Previous disclosures for continuing topics must also be updated to reflect changes in a member's situation since the form was last completed.
Prior to each meeting or to new member appointment, all disclosures are reviewed by the Task Force Chairs according to the criteria specified in the USPSTF Procedure Manual and determined to be either Level 1, 2, or 3. The Task Force Chairs determine the final action on the member's eligibility to participate on a specific topic based on the nature and significance of the potential conflict.

- Level 1 disclosures include nonfinancial disclosures that would not affect the judgment of a Task Force member. These disclosures do not require any action.
- Level 2 disclosures include financial disclosures of $1,000 or less and nonfinancial disclosures that are relevant to a topic but not anticipated to affect the judgment of the Task Force member for that topic. These disclosures are announced at the Task Force meeting, but do not limit the Task Force member's participation in the topic process.
- Level 3 disclosures include financial disclosures of a larger amount and significant nonfinancial disclosures that may affect the Task Force member's view on the topic. Actions for Level 3 disclosures vary according to the nature of the conflict, and may include preventing the member from serving as lead of a topic or on the workgroup of a topic, preventing the member from serving as a primary spokesperson for a topic, or preventing the member from taking part in all topic activities. As all new Task Force members are reviewed for conflicts prior to joining the Task Force, Level 3 disclosures are extremely rare.

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: PHS does not separately grade the strength of the recommendation; USPSTF Grade A

USPSTF 2015 Recommendation Statement

Grade A

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Recommendations from the USPSTF *are considered the gold standard for clinical preventive services.* The USPSTF is an independent panel of nonfederal experts in prevention and evidence-based medicine. The Task Force carefully assesses the evidence and makes recommendations about preventive services such as screening tests, counseling services, or preventive medications that are provided in clinical settings, and are intended to prevent disease or improve health outcomes from heart disease, cancer, infectious diseases, and other conditions and events that affect the health of children, adolescents, adults, older adults, and pregnant women.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency: High

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

0028_and_3185_Evidence_MSF5.0_Updated_for_2016_Submission-636162891174792000.doc

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

This measure is intended to promote adult tobacco screening and tobacco cessation interventions for those who use tobacco products. There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.

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 (4b) under Usability and Use.*

2014 Physician Quality Reporting System (PQRS) Experience Report

2014 is the most recent year for which PQRS Experience Report measure data are available. The average performance rates on Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention over the last several years are as follows:

- 2011: 81.6%
- 2012: 84.1%
- 2013: 89.7%
- 2014: 88.9%

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. However, beginning in 2015, the program will impose payment penalties for non-participants based on 2013 performance. For 2014, only 21.7% of eligible professionals reported on the measure.

As a result, performance rates may not be nationally representative.

Reference: Center for Medicare and Medicaid Services. 2014 Reporting Experience Including Trends. Available:

https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

2015 PQRS EHR Performance Rate:
Mean: 76.38%
Minimum: 0.00%
Maximum: 100.00%

Decile	Result %
1	27.84%
2	62.07%
3	76.50%
4	83.44%
5	87.91%
6	91.11%
7	93.69%
8	96.03%
9	98.41%
10	100.00%

Report Title: PQRS Ad Hoc Analysis PQ3783, 2015 PQRS Measure Data for PCPI
Report includes 2015 Part B Claims Data for services rendered between January 1, 2015 and December 31, 2015 and processed through February 2016 TAP.
Report also includes PQRS Final Action Registry data and 2015 PQRS Final Action EHR data.

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.

A number of studies have documented low rates of tobacco use screening and cessation intervention during primary care and other office/outpatient visits, missing key opportunities for intervention.

A 2012 Morbidity and Mortality Weekly Report (MMWR) summarized data from 2005-2008 National Ambulatory Medical Care Survey (NAMCS) and the National Health Interview Survey (NHIS) to determine progress toward Healthy People 2020 objectives calling for increased screening, cessation counseling and cessation success. The following key findings were reported:

- During the study period, adults aged 18 years and older made an estimated annual average of approximately 771 million outpatient visits (an estimated total of 3.08 billion visits during 2005–2008 combined) to office-based physicians.
- Tobacco use screening occurred during the majority of adult visits to outpatient physician offices (62.7%)
- Of the visits that included tobacco use screening, 17.6% (340 million visits) were made by current tobacco users.
- Among patients who were identified as current tobacco users, only 20.9% received tobacco cessation counseling and 7.6% received tobacco cessation medication
- Patients who visited their primary care physician were more likely to receive tobacco screening (66.6% of visits) than patients who visited a physician who was not their primary care physician (61.6% of visits). Screening also varied by physician specialty. Patients visiting general or family practitioners (66.4%) and obstetricians/gynecologists (69.6%) were more likely to receive screening than patients who visited physicians in other specialties (58.2%), excluding internal medicine, cardiovascular disease, and psychiatry. (1)

Given that hospital outpatient visits account for approximately 1 in 10 outpatient visits, Jamal and colleagues sought to assess the rates of tobacco use screening and cessation assistance offered to US adults during their hospital outpatient clinic visits analyzing data from the 2005–2010 NAMCS. The following key findings were reported:

- During the study period, adults aged 18 years or older made, on average, 71.8 million hospital outpatient visits annually to hospital outpatient physicians or an estimated 431 million visits from 2005 through 2010 combined.
- On average, 45.2 million (63.0%) hospital outpatient visits included tobacco use screening each year.
- Of the visits that included tobacco use screening, 25.7% (11.6 million annual average visits) were made by current tobacco users.

- Among patients who screened positive for current tobacco use, 24.5% (or an estimated 17.1 million visits) received any cessation assistance, including tobacco counseling, a prescription or order for a cessation medication at the visit, or both.
- Patients who made visits to general medicine clinics (67.1%) were more likely to receive tobacco use screening than those who made visits to surgical clinics (55.7%) or clinics with other specialties (45.2%), excluding obstetrics and gynecology (62.8%) and substance abuse clinics (68.3%). (2)

Citations:

1. Jamal A1, Dube SR, Malarcher AM, Shaw L, Engstrom MC; Centers for Disease Control and Prevention (CDC). Tobacco use screening and counseling during physician office visits among adults--National Ambulatory Medical Care Survey and National Health Interview Survey, United States, 2005-2009. MMWR Suppl. 2012 Jun 15;61(2):38-45.
2. Jamal A, Dube SR, King BA. Tobacco Use Screening and Counseling During Hospital Outpatient Visits Among US Adults, 2005–2010. Prev Chronic Dis 2015;12:140529.

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 (4b) 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**

The MMWR noted that rates of tobacco screening and intervention varied by patients' race, age and insurance status. Overall, patients classified as non-Hispanic whites were more likely to receive counseling than Hispanic patients (64.1 versus 57.8%). Among current tobacco users, younger patients (aged 25 to 44 years) reported receiving less counseling (17.9%) than patients aged 45 to 64 years (22.7%). Patients with workers' compensation, and those whose insurance status was unknown were less likely to receive counseling than those with private insurance, self-payers, Medicaid, and Medicare patients.

Similar racial/ethnic disparities were reported for hospital outpatient visits. Tobacco use screening varied by patient's race/ethnicity - visits made by Hispanics (55.4%) were less likely to receive tobacco use screening than those by non-Hispanic whites (65.1%). For tobacco users, cessation assistance was higher for visits made by those with Medicaid/SCHIP (27.6%) than those with private insurance (21.8%) or Medicare (21.4%). Patients living in a high poverty zone were more likely to receive cessation than those living in a low poverty zone. (2)

1. Jamal A1, Dube SR, Malarcher AM, Shaw L, Engstrom MC; Centers for Disease Control and Prevention (CDC). Tobacco use screening and counseling during physician office visits among adults--National Ambulatory Medical Care Survey and National Health Interview Survey, United States, 2005-2009. MMWR Suppl. 2012 Jun 15;61(2):38-45.
2. Jamal A, Dube SR, King BA. Tobacco Use Screening and Counseling During Hospital Outpatient Visits Among US Adults, 2005–2010. Prev Chronic Dis 2015;12:140529.

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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 : 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.)*
The measure specifications are included as an attachment with this submission. Additional measure details may be found at: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/eCQM_Library.html Value sets at: https://vsac.nlm.nih.

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_CMS138v5_NQF0028_PREV_Tobacco-636162883268556000.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:** EP_CMS138v5_NQF0028_ValueSets_20160401.xlsx

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.
Supporting guidelines and coding value sets included in the measure are reviewed on an annual basis. The updated recommendation from USPSTF published in 2015 resulted in updated clinical recommendation statements and guidance regarding the intentional omission of electronic nicotine delivery systems (ENDS) from the measure. Additional limited changes have been incorporated into the technical specifications to adhere to current eCQM industry standards while preserving the original measure intent.

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 who were screened for tobacco use at least once within 24 months AND who received tobacco cessation intervention if identified as a tobacco user

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 least once during the 24 month period

Definitions:
Tobacco Use – Includes any type of tobacco
Tobacco Cessation Intervention – Includes brief counseling (3 minutes or less), and/or pharmacotherapy

For EHR:
HQMF eMeasure developed and is attached to this submission in fields S.2a and S.2b.
NUMERATOR GUIDANCE:
If a patient uses any type of tobacco (ie, smokes or uses smokeless tobacco), the expectation is that they should receive tobacco cessation intervention: either counseling and/or pharmacotherapy.

If tobacco use status of a patient is unknown, the patient does not meet the screening component required to be counted in the numerator and should be considered a measure failure. Instances where tobacco use status of "unknown" is recorded include: 1) the patient was not screened; or 2) the patient was screened and the patient (or caregiver) was unable to provide a definitive answer. If the patient does not meet the screening component of the numerator but has an allowable medical exception, then the patient should be removed from the denominator of the measure and reported as a valid exception.

As noted above in a recommendation statement from the USPSTF, the current evidence is insufficient to recommend electronic nicotine delivery systems (ENDS) including electronic cigarettes for tobacco cessation. Additionally, ENDS are not currently classified as tobacco in the recent evidence review to support the update of the USPSTF recommendation given that the devices do not burn or use tobacco leaves. In light of the current lack of evidence, the measure does not currently capture e-cigarette usage as either tobacco use or a cessation aid.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 18 years and older seen for at least two visits or at least one preventive visit during the measurement period

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

For EHR:
HQMF eMeasure 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)*
Documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason)

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.)*
Time Period for Data Collection: At least once during the 24 month period

Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For measure Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention, exceptions may include documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason). Where examples of exceptions are included in the measure language, value sets for these

examples are developed and included in the eMeasure. Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

For EHR:
HQMF eMeasure developed and is attached to this submission in fields S.2a and S.2b.
DENOMINATOR EXCEPTION GUIDANCE:
The medical reason exception only applies to the screening data element of the measure; once a patient has been screened, there are no allowable medical reason exceptions for not providing the intervention.

If a patient has a diagnosis of limited life expectancy, that patient has a valid denominator exception for not being screened for tobacco use or for not receiving tobacco use cessation intervention (counseling and/or pharmacotherapy) if identified as a tobacco user.

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 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
4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason). If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, 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 a 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.)*
IF a PRO-PM, 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 Record (Only)

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 is collected.)*
IF a PRO-PM, identify the specific PROM(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)*
Behavioral Health : Outpatient, Clinician Office/Clinic, Home Health, Other
If other: Occupational therapy evaluation, speech and hearing evaluation, ophthalmological services visit

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. The measure is not a composite.

2. Validity – See attached Measure Testing Submission Form
NQF3185_TobaccoTesting_Attachement_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. (Do not remove prior testing information – include date of new information in red.)
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. (Do not remove prior testing information – include date of new information in red.)
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE:

These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 3185
Measure Title: Preventive Care & Screening: Tobacco Use: Screening & Cessation Intervention
Date of Submission: 12/2/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.

- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have

differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

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

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

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

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

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

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

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

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

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

The data source is 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? Click here to enter date range

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.26***)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice

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

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

The data source is EHR data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS). The data are for the time period January 2015 through December 2015 and cover the entire United States.

The total number of physicians reporting on this measure, via the EHR option, in 2015, is 42,902. Of those, 39,291 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 20,702,162 quality events. For this measure, 91.6 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 524.5 for the remaining 20,610,094 events. The range of quality reporting events for 39,291 physicians included is from 19,263 to 10. The average number of quality reporting events for the remaining 8.4 percent of physicians that aren't included is 3.84.

There were 20,610,094 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 and remained after exceptions were removed.

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*)

There were 20,610,094 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 and remained after exceptions were removed

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.

The same data sample was used for reliability testing and exceptions analysis.

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

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*)
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, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

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*)

EHR – Signal to Noise Ratio analysis (PQRS)

This measure has 0.81 reliability when evaluated at the minimum number of quality reporting events and 0.99 reliability when evaluated at the average number of quality events.

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?*)

Reliability at the minimum level of quality reporting events is high. Reliability at the average number of quality events is very high.

2b2. VALIDITY TESTING

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

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

☒ **Performance measure score**

☐ **Empirical validity testing**

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

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

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

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.

Scale 1-5, where 1= Strongly Disagree; 2= Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

The expert panel included 10 members. Panel members were comprised of the newly convened PCPI Preventive Care Technical Expert Panel who did not participate in the original workgroup.

The list of expert panel members is as follows:

Sandra Dunbar, PHD, RN
Peter Briss, MD, MPH
Yngve Falck, MD
Susan Friedman, MD, MPH
Marc Ghany, MD
Ashley Halle, OTD, OTR/L
Selena Hariharan, MD
Lori Karan, MD
Andrew J Saxon, MD
John Wong, MD

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's ICD-10 conversion approach was used to identify ICD-10 codes for this measure. 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 two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, 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.

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

Frequency Distribution of Ratings

1 – 1 responses (Strongly Disagree)
2 – 0 responses (Disagree)
3 – 3 responses (Neither Agree nor Disagree)
4 – 4 responses (Agree)
5 – 2 responses (Strongly Agree)

The results of the expert panel rating of the validity statement were as follows: N = 10; Mean rating = 3.6 and 60.0% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)
Based on the mean rating by the expert panel, this measure is valid as specified.

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

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

EHR – PQRS Exceptions Analysis (PQRS)

Exceptions included documentation of medical reason for not screening for tobacco use. Exceptions were analyzed for frequency and variability across providers.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

EHR – PQRS Exceptions Analysis (PQRS)

Amongst the 39,291 physicians with the minimum (10) number of quality reporting events, there were a total of 92,068 exceptions reported. The average number of exceptions per physician in this sample is 2.3. The overall exception rate is 0.4%.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Exceptions are necessary to account for those situations when it is not medically appropriate for a patient to have tobacco screening. Exceptions are discretionary and the methodology used for measure exception categories are not uniformly relevant across all measures; for this measure, there is a clear rationale to permit an exception for medical reasons. Rather than specifying an exhaustive list of explicit reasons for exception for this measure, the measure developer relies on clinicians to link the exception with a specific medical reason for the decision to screen for tobacco use.

Some have indicated concerns with exception reporting including the potential for physicians to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by physicians, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008).

Although this methodology does not require the external reporting of more detailed exception data, the measure developer recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Without exceptions, the performance rate would not accurately reflect the true performance of that physician. This would result in an increase in performance failures and false negatives. The additional value of increased data collection of capturing an exception greatly outweighs the reporting burden.

References:

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. New Engl J Med. 2008; 359: 274-84.

Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. Ann Intern Med. 2011;154:227-234

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

EHR –Exceptions Analysis (PQRS)

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

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:

EHR – PQRS Exceptions Analysis (PQRS)

Exceptions included documentation of medical reason(s) for not screening for tobacco use. Exceptions were analyzed for frequency and variability across providers.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

EHR – PQRS Exceptions Analysis (PQRS)

Amongst the 39,291 physicians with the minimum (10) number of quality reporting events, there were a total of 92,068 exceptions reported. The average number of exceptions per physician in this sample is 2.3. The overall exception rate is 0.4%.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

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

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

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

2b4.1.1 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

2b4.2. 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

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not applicable

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

Not applicable

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable

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

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

If stratified, skip to 2b4.9

Not applicable

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Not applicable

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable

2b4.9. Results of Risk Stratification Analysis:

Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not applicable

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

EHR –Signal to Noise Ratio analysis (PQRS)

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

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

EHR – Signal to Noise Ratio analysis (PQRS)

Based on the sample of 39,291 included physicians, the mean performance rate is 0.76 the median performance rate is 0.87 and the mode is 0. The standard deviation is 0.27. The range of the performance rate is 1, with a minimum rate of 0 and a maximum rate of 1. The interquartile range is 0.23 (0.71 – 0.94).

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The range of performance from 0.71 to 0.94 suggests there's clinically meaningful variation across physicians' performance.

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
This test was not performed for this measure.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
This test was not performed for this measure.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
This test was not performed for this measure.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
Data are not available to complete this testing.

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

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: NQF3185_FeasiblityReport.pdf,Tobacco_Feasibility_Scorecard_v1.0_PCPI_SITE1.xlsx

3c. Data Collection Strategy

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

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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, eg, 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(R) Foundation (PCPI[R]) or the American Medical Association (AMA). Neither the American Medical Association (AMA), nor the AMA-convened Physician Consortium for Performance Improvement(R) (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 Physician Quality Reporting System (PQRS) http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/MeasuresCodes.html Public Health/Disease Surveillance Million Hearts Initiative http://millionhearts.hhs.gov Payment Program Meaningful Use Stage 2 (EHR Incentive Program) http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/eCQM_Library.html

4a.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 Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)
PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013.
Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented
a phased approach to public reporting performance information on the Physician Compare Web site.

CMS has implemented a phased approach to publicly reporting performance information on the Physician Compare Web Site. Beginning with PQRS 2014 reporting, this measure is one of 14 group practice PQRS measures reported via the Web interface that are currently available for public reporting. This measure is also one of 6 individual EP PQRS measures reported via claims that are currently available for public reporting. CMS also announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in 2017. Beginning in 2017, the Merit-based Incentive Payment System (MIPS) consolidates PQRS and other existing quality reporting programs. This measure has been finalized as an individual quality measure available for MIPS reporting in 2017.

Million Hearts
Million Hearts™ is a national initiative to prevent 1 million heart attacks and strokes in the U.S. over the next 5 years. Launched by the Department of Health and Human Services (HHS) in September 2011, it aligns existing efforts, as well as creates new programs, to improve health across communities and help Americans live longer, more productive lives. The Centers for Disease Control and Prevention (CDC) and Centers for Medicare & Medicaid Services (CMS), co-leaders of Million Hearts™ within HHS, are working alongside other federal agencies and private-sector organizations to make a long-lasting impact against cardiovascular disease.

The Million Hearts® Clinical Quality Measures (CQM) Dashboard is designed to display quality reporting measures focused on the Million Hearts®ABCS (Aspirin when appropriate, Blood pressure control, Cholesterol management, and Smoking cessation) and is based on information from the following available data systems, where possible.

HRSA UDS - Health Resources and Services Administration Uniform Data System
NCQA HEDIS - National Committee for Quality Assurance Healthcare Effectiveness Data and Information Set
CMS PQRS - Centers for Medicare & Medicaid Services Physician Quality Reporting System

Meaningful Use Stage 2 (EHR Incentive Program) – Sponsored by the Centers for Medicare and Medicaid Services (CMS)
The Medicare and Medicaid EHR Incentive Programs provide incentive payments to eligible professionals, eligible hospitals, and critical access hospitals (CAHs) as they adopt, implement, upgrade or demonstrate meaningful use of certified EHR technology. Eligibility for incentive payments for the “meaningful use” of certified EHR technology is established if all program requirements are
met, including successful implementation and reporting of program measures, which include this measure, to demonstrate meaningful use of EHR technology.

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

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

4b. **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.
Although the PQRS program has demonstrated increasing performance rates over time which would indicate progress on improvement, it's important to note that the percentage of eligible professional reporting on PQRS measures overall and on this measure, in particular, continues to grow but remains low. In 2014, for example, only 21.7% of eligible professionals reported on the measure. As a result, performance rates may not be nationally representative.

Additionally, 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.

4c. Unintended Consequences
The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **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 measure.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
We are not yet aware of any unexpected benefits related to this measurement.

4d1.1. **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, 32 individuals from a diverse group of specialties including family medicine, internal medicine, geriatric medicine, gastroenterology, general surgery, nursing, and psychology participated on the measure development work group.

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 works 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.

4d1.2. **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 4d1.1 above.

4d2.1. **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.

4d2.2. **Summarize the feedback obtained from those being measured.**
The majority of comments received during public comment were supportive and approving of the broad nature of the measure, its potential for public health impact and patient outcomes. There were some specific comments requesting consideration of a lower age range for the measure and adding a medical reason exception for patients with limited life expectancy.

The majority of feedback from implementers seeks to have the PCPI clarify what qualifies and does not qualify as meeting the measure. More recently, many implementers wanted to understand how the measure addresses electronic nicotine delivery systems (ENDS).

4d2.3. **Summarize the feedback obtained from other users**

See summary in 4d2.2 above.

4d.3. **Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
At the time of original development, the expert work group decided not to adjust the age range as it was developed to align with the USPSTF's recommendation for adults. The latter comment regarding the medical reason exception was incorporated into the final version of the measure.

As a result of implementation feedback, a brief definition of cessation intervention has been added to the measure. Guidance has been provided to explain the omission of ENDS from the measure and the rationale for doing so.

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)**
0027 : Medical Assistance With Smoking and Tobacco Use Cessation
1651 : TOB-1 Tobacco Use Screening
1654 : TOB - 2 Tobacco Use Treatment Provided or Offered and the subset measure TOB-2a Tobacco Use Treatment
1656 : TOB-3 Tobacco Use Treatment Provided or Offered at Discharge and the subset measure TOB-3a Tobacco Use Treatment at Discharge
2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
2803 : Tobacco Use and Help with Quitting Among Adolescents

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?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Related measures have differing target populations and/or levels of measurement from the PCPI's Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention measure. 3185 is the emeasure version of 0028 and focuses on routine tobacco screening for all adults and tobacco cessation interventions for those who use tobacco products and is intended to assess clinician level performance towards these objectives. The cessation intervention required by the PCPI measure includes brief counseling and/or pharmacotherapy in light of the strong support for these interventions in the guidelines and the feasibility of implementing these practices as part of routine care. Measure 0027 is a patient survey measure assessing health

plan performance and includes one additional component of the cessation intervention beyond our measure (ie, discussion of methods or strategies other than medication). Measures 1651, 1654 and 1656 assess hospital level performance at providing tobacco use and treatment to patients being discharged from hospitals. Measure 2803 is focused on assessing clinical level performance on tobacco cessation counseling among adolescents. Finally, measure 2600 represents an adaptation of the PCPI measure and is limited to a subset of the population of patients with serious mental illness.

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.)
No competing measures.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Tobacco_Feasibility_Scorecard_v1.0_PCPI_SITE2.xlsx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** PCPI Foundation
Co.2 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-464-5524-
Co.3 **Measure Developer if different from Measure Steward:** PCPI Foundation
Co.4 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-464-5524-

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.
Gail M. Amundson, MD, FACP (internal medicine/geriatrics)
Joel V. Brill MD, AGAF, FASGE, FACG (gastroenterology)
Steven B. Clauser, PhD
Will Evans, DC, Phd, CHES (chiropractic)
Ellen Giarelli, EdD, RN, CRNP (nurse practitioner)
Amy L. Halverson, MD, FACS (colon & rectal surgery)
Alex Hathaway, MD, MPH, FACPM
Charles M. Helms, MD, Phd (infectious disease)
Kay Jewell, MD, ABHM (internal medicine/geriatrics)
Daniel Kivlahan, PhD (psychology)
Paul Knechtges, MD (radiology)
George M. Lange, MD, FACP (internal medicine/geriatrics)
Trudy Mallinson, PhD, OTR/L/NZROT (occupational therapy)
Elizabeth McFarland, MD (radiology)
Jacqueline W. Miller, MD, FACS (general surgery)
Adrienne Mims, MD, MPH (geriatric medicine)
Sylvia Moore PhD, RD, FADA (dietetics)

G. Timothy Petito, OD, FAAO (optometry)
Rita F. Redberg, MD, MSc, FACC (cardiology)
Barbara Resnick, PhD, CRNP (nurse practitioner)
Sam JW Romeo, MD, MBA (family practice)
Carol Saffold, MD (obstetrics & gynecology)
Robert A. Schmidt, MD (radiology)
Samina Shahabbudin, MD (emergency medicine)
James K. Sheffield, MD (health plan representative)
Arthur D. Snow, MD, CMD (family medicine/geriatrics)
Richard J. Snow, DO, MPH
Brooke Steele, MD
Brian Svazas, MD, MPH, FACOEM, FACPM (preventive medicine)
David J. Weber, MD, MPH (infectious disease)
Deanna R. Willis, MD, MBA, FAAFP (family medicine)
Charles M. Yarborough, III, MD, MPH (occupational medicine)
PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be 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. 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. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2001

Ad.3 **Month and Year of most recent revision:** 11, 2015

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?** 11, 2016

Ad.7 **Disclaimers:** See copyright statement above.

Ad.8 **Additional Information/Comments:** Due to file count restrictions, the feasibility scorecard for site 2 is included in the appendix (A1).

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 #: 3205
Measure Title: Medication Continuation Following Inpatient Psychiatric Discharge
Measure Steward: Centers for Medicare & Medicaid Services, Contracting Officer's Representative (COR)
Brief Description of Measure: This measure assesses whether psychiatric patients admitted to an inpatient psychiatric facility (IPF) for major depressive disorder (MDD), schizophrenia, or bipolar disorder filled a prescription for evidence-based medication within 2 days prior to discharge and 30 days post-discharge. The performance period for the measure is two years.
Developer Rationale: The aim of the proposed measure is to address gaps in continuity of pharmaceutical treatment during the transition from inpatient care to outpatient care. Pharmacotherapy is the primary form of treatment for most patients discharged from an inpatient psychiatric facility (IPF) for major depressive disorder (MDD), schizophrenia, or bipolar disorder. The measure focuses on medication continuation because it is an essential step in medication adherence.

Medication continuation is particularly important in the psychiatric patient population because psychotropic medication discontinuation can have a range of adverse effects, from mild withdrawal to life-threatening autonomic instability and psychiatric decompensation (Ward & Schwartz, 2013). Patients with MDD who do not remain on prescribed medication are more likely to have negative health outcomes, such as relapse and readmission, decreased quality of life, and increased healthcare costs. If untreated, MDD can contribute to or worsen chronic medical disorders (Geddes et al., 2003; Glue et al., 2010). The literature shows that among patients with schizophrenia, those who were "good compliers" according to the Medication Adherence Rating Scale had better outcomes in terms of rehospitalization rates and medication maintenance (Jaeger et al., 2012). Among patients with bipolar disorder, medication adherence was significantly associated with reduction in manic symptoms (Sylvia et al., 2013), while non-adherence was associated with increased suicide risk (OR 10.8, CI 1.57–74.4; Gonzalez-Pinto et al., 2006).

Current facility-level performance indicates that there is a clear quality gap. Using 2013–2014 Medicare claims data, we found that there is about a 22 percentage point difference between the 10th and 90th percentiles (66.7%-88.3%) and a median score of 79.6%. By calculating the facility-level rates of medication continuation in Medicare FFS claims data, this measure can provide valuable information on areas where care transitions to the outpatient setting can be improved.

*Geddes, J. R., Carney, S. M., Davies, C., Furukawa, T. A., Kupfer, D. J., Frank, E., & Goodwin, G. M. (2003). Relapse prevention with antidepressant drug treatment in depressive disorders: A systematic review. The Lancet, 361(9358), 653–661. doi:10.1016/s0140-6736(03)12599-8
*Glue, P., Donovan, M. R., Kolluri, S., & Emir, B. (2010). Meta-analysis of relapse prevention antidepressant trials in depressive disorders. Australian and New Zealand Journal of Psychiatry, 44(8), 697-705. doi: 10.3109/00048671003705441

*Gonzalez-Pinto, A., Mosquera, F., Alonso, M., López, P., Ramírez, F., Vieta, E., & Baldessarini, R. J. (2006). Suicidal risk in bipolar I disorder patients and adherence to long-term lithium treatment. Bipolar Disorders, 8(5p2), 618–624. doi:10.1111/j.1399-5618.2006.00368.x
*Jaeger, S., Pfiffner, C., Weiser, P., Kilian, R., Becker, T., Langle, G., . . . Steinert, T. (2012). Adherence styles of schizophrenia patients identified by a latent class analysis of the Medication Adherence Rating Scale (MARS): A six-month follow-up study. Psychiatry Research, 200(2-3), 83-88. doi: 10.1016/j.psychres.2012.03.033
*Sylvia, L. G., Hay, A., Ostacher, M. J., Miklowitz, D. J., Nierenberg, A. A., Thase, M. E., . . . Perlis, R. H. (2013). Association between therapeutic alliance, care satisfaction, and pharmacological adherence in bipolar disorder. Journal of Clinical Psychopharmacology, 33(3), 343-350. doi: 10.1097/JCP.0b013e3182900c6f
*Ward, M., & Schwartz, A. (2013). Challenges in pharmacologic management of the hospitalized patient with psychiatric comorbidity. Journal of Hospital Medicine, 8(9), 523–529. doi:10.1002/jhm.2059

Numerator Statement: The numerator for this measure includes:
1. Discharges with a principal diagnosis of MDD in the denominator population for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge
2. Discharges with a principal diagnosis of schizophrenia in the denominator population for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge
3. Discharges with a principal diagnosis of bipolar disorder in the denominator population for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge

Denominator Statement: The target population for this measure is Medicare fee-for-service (FFS) beneficiaries with Part D coverage aged 18 years and older discharged from an inpatient psychiatric facility with a principal diagnosis of MDD, schizophrenia, or bipolar disorder.

Denominator Exclusions: The denominator for this measure excludes discharged patients who:
1. Received Electroconvulsive Therapy (ECT) during the inpatient stay or follow-up period.
2. Received Transcranial Magnetic Stimulation (TMS) during the inpatient stay or follow-up period.
3. Were pregnant during the inpatient stay.
4. Had a secondary diagnosis of delirium.
5. Had a principal diagnosis of schizophrenia with a secondary diagnosis of dementia.

Measure Type: Process
Data Source: Claims (Only)
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.

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.
- Major depressive disorder (MDD)
  - APA 2010 Guidelines support the use of antidepressant medications for acute and maintenance treatment (except with ECT). **(Grade I Recommendation: substantial clinical confidence)**
  - VA/DoD 2016 Guidelines support the use of antidepressant medications for at least 6 months after remission **(Grade A Recommendation: good evidence, benefits substantially outweigh harm).**
- Schizophrenia
  - APA 2010 Guidelines support use of medications in acute phase **(Grade I Recommendation: substantial clinical confidence)**, for long-acting injectable medications for those with recurrent relapses **(Grade II Recommendation: moderate clinical confidence)**, and for continued medication for at least 6 months if improvement is noted **(Grade I Recommendation: substantial clinical confidence)**
  - The developer noted a new study since the guideline's release comparing longer-term effects and usefulness of a range of antipsychotics, supporting the inclusion of both typical and atypical antipsychotics in this measure.
- Bipolar Disorder
  - APA 2002 Guidelines support use of medications, describing a variety of options/medication choices depending on the situation **(Grade I and Grade II Recommendations: substantial or moderate clinical confidence)**, especially for continuation of medication after remission **(Grade I Recommendation: substantial clinical confidence)**.
  - VA/DoD 2010 Guidelines support the use of various medications **(Grade A, B, and I Recommendations: A-good evidence and benefits substantially outweigh harms; B-fair evidence and benefits outweigh harms; I-evidence on effectiveness is lacking or poor quality or conflicting and balance of benefits and harms cannot be determined)** and the use of medications for continued maintenance after an initial acute manic episode, for at least 6 months **(Grade A Recommendation)**.
    - Note that the Grade I recommendations mostly apply to medications to be used as a secondary choice.
- NOTE: The developer provides guidelines which emphasize the need for continued use of medications, but the evidence described largely focuses on the efficacy or relative advantage of individual medications and not on the timeliness of their use (as is the focus of this measure).
- The VA/DoD guidelines provide some insight to the quality of the studies, but overall the quality of the evidence has not been presented.

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *Is more evidence needed that the timeliness of filling prescriptions leads to better outcomes?*

Guidance from the Evidence Algorithm

Process measure supported by systematic review (Box 3) → QQC not provided (Box 4) → guideline are mostly supported by strong recommendations (Box 6) →Moderate

The highest possible rating is MODERATE.

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

1b. Gap in Care/Opportunity for Improvement a**nd 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 results at the facility level using 2013-2014 Medicare claims data across 1,694 facilities.
- Median performance by diagnosis:
  - MDD – 77.1%
  - Schizophrenia – 81.5%
  - Bipolar disorder – 80.0%
- The overall distribution of scores is as follows:

Mean	Standard Deviation	Min	10th Percentile	Median	90th Percentile	Max
78.0%	11.1%	0.0%	66.7%	79.6%	88.3%	100%

Disparities

- The developer described that they looked at disparities by performing an analysis to determine if the difference between a specific population and a reference group is statistically significant ($p<0.05$ on a two-tailed test), and whether the relative difference between the population and the reference group is at least 10%.
- The developer notes their analysis found
  - Black patients have significantly worse rates of medication continuation than the reference group.
  - Dually-enrolled patients have significantly better rates of medication continuation than patients enrolled in only Medicare.
  - There are no differences in performance among age groups.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Do you agree the analysis of disparities is reasonable?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

1a. Evidence to Support Measure Focus

Comments:

**Fairly strong evidence to support measure
**logic model. Evidence shows discontinuing psychiatric medications can lead to relapse.
**The relationship between filling prescribed medications and patient outcomes is identified and supported. More evidence is not needed that the timeliness of filling prescriptions will lead to better outcomes.
**The logic model makes sense on why this metric is being proposed, however the literature provided for evidence to support metric listed need for long-term continuation of medications, types of medications for each identified metric disease state, but there was no study supporting post-hospital discharge drug use/initiation/continuation. However, from my own work at one inpatient psychiatric hospital in Detroit, MI, continuation of medications in the ambulatory setting is a big concern. We have found a difference based upon where patients go in the ambulatory setting and I didn't see that defined in this metric.

1b. Performance Gap

Comments:

**There is somewhat a performance gap and there is disparity demonstrated.

**Not a very large performance gap but meaningful. Median was 80%. Lower quartile was 74% and highest quartile was 84%.
**Performance data is provided and it demonstrates a gap in care. Population subgroup data was provided.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

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

Data source(s): Claims (only) [Medicare administrative data from Parts A, B, and D claims.]

Specifications:

- The level of analysis is at the facility level.
- A higher score indicates better quality
- Patients included in the measure denominator include Medicare fee-for-service (FFS) beneficiaries with Part D coverage aged 18 years and older discharged from an inpatient psychiatric facility with a principal diagnosis of MDD, schizophrenia, or bipolar disorder.
    - ICD-9-CM and ICD-10-CM codes are provided to identify MDD, schizophrenia, and bipolar disorder.
- Patients included in the measure numerator include discharges with a principal diagnosis of MDD, schizophrenia, or bipolar disorder for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge.
    - The developer provides a list of medications for MDD, schizophrenia, and bipolar disorder.
- Exclusions include
    - Electroconvulsive therapy (ECT) or transcranial magnetic stimulation (TMS) during inpatient stay or follow-up period.
    - Pregnancy during inpatient stay
    - Secondary diagnosis of delirium
    - Principal diagnosis of schizophrenia with secondary diagnosis of dementia
- A detailed calculation algorithm is provided.
- The measure is not risk adjusted.
- The developer notes that a minimum denominator of 75 discharges is needed to achieve adequate reliability.

Questions for the Committee:

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

2a2. Reliability Testing, Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data used for testing included Medicare files for all inpatient psychiatric facility (IPF) discharges that occurred between January 1, 2013 and December 31, 2014. This included 380,861 discharges from 1,694 IPFs.
- The developer provides distribution of discharges by diagnosis.
- Developers conducted a signal-to-noise analysis which is an appropriate method for testing reliability.
- A signal-to-noise analysis quantifies the amount of variation in a performance measure that is due to true differences between providers (i.e., signal) as opposed to measurement error (i.e., noise). Results will vary based on the amount of variation between health plans (or states) and the number of patients treated by each health plan (or state). A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in health plan (or state) performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Results of reliability testing

- The developer notes that a minimum denominator of 75 discharges is needed to attain an overall reliability score of at least 0.7. With a 2-year measurement period, 70% of IPFs had enough discharges for public reporting. They noted that removal of facilities with less than 75 discharges in the measurement period did not have an appreciable impact on the distribution of scores.

IPF Reliability and Assessment of Adequacy for Tests Conducted

	Minimum Denominator	# of IPFs N=1,694 (%)	Mean Rate (%) of IPFs	Reliability Score
Overall	75	1,184 (69.9)	78.0	0.77

Comparison of IPF Measure Score Distribution by Denominator Minimum

	# IPFs	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
Overall	1,694	78.0	11.1	0.0	66.7	73.6	79.6	84.4	88.3	100.0
Denominator ≥ 75	1,184	78.0	7.9	21.1	68.3	73.9	79.1	83.4	86.5	98.5

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm [Algorithm guidance]

Submitted specifications are precise, unambiguous and complete (Box 1) → Empirical reliability analysis conducted with measure as specified, except that reliability score is limited to IPFs with at least 75 discharges (Box 2) → Reliability testing was conducted with computed performance measure score (Box 4) → Method was appropriate to assess variability in performance at measured entity level (Box 5) → Level of certainty that measure is reliable (Box 6) → Moderate

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Specification not completely consistent with evidence The evidence presented focused primarily on the efficacy of medications and not the timeliness of filling prescriptions, but the guidelines do speak to the importance of medication continuation.

Question for the Committee:

o *Are the specifications consistent with the evidence?*

2b2. Validity testing

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

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☒ **Both**

Method of validity testing of the measure score:

☒ **Face validity ~~only~~**

Validity testing method:

- Data element validity testing
  - Two psychiatrists reviewed 150 patient records.
    - Clinicians' recorded assessments of principal discharge diagnosis were compared to claims.
    - Positive predictive value (PPV) was calculated using the clinical assessment from the medical record as the gold standard. Note: a high PPV indicates high probability that a claim for a specific condition correctly predicts the diagnosis at discharge in the medical record.
  - Additionally, abstractors at 7 sites indicated whether a prescription was provided at discharge, and if not, to provide a rationale in order to determine if additional exclusions were needed.
    - Data on provision of at least one prescription was compared to claims data.
    - PPV was calculated , indicating that most patients who filled a prescription in the follow-up period also received a prescription at discharge.
  - Abstractors at the 7 sites also recorded if the medical record indicated medications were dispensed to the patient free at discharge (since those would not be reflected in claims data).
  - Ten percent of all abstraction cases were reviewed by both clinicians.
- Measure score validity testing
  - Measure scores were compared to three related measures:
    - Follow-Up After Hospitalization (7-Day)
    - Follow-Up After Hospitalization (30-Day)
    - IPF All-Cause Unplanned Readmission Measure
  - The developer hypothesized the first two measures would be positively correlated with the medication continuation scores, as they all reflect care coordination. The developer hypothesized that the medication continuation score would be negatively correlated with the all-cause unplanned readmission score, "because readmissions may indicate a lack of care coordination."
- Face validity
  - Face validity of the measure score was assessed by a technical expert panel. Members were asked if they agreed if the performance rating as specified accurately represents facility-level rates of medication continuation.

Validity testing results:

- Data element validity testing
  - PPV of claims data was 97%. (MDD – 98%; schizophrenia – 98%; bipolar disorder -96%).
  - For the medical record review, 92% of cases were prescribed medication at discharge ; PPV was 96%.
  - Few discharges included provision medications at discharge.

- Measure score validity testing
  - The developer appears to have done a Pearson correlation, which measures the degree of association between two quantitative variables. For the social sciences, scores of 0.37 or larger are considered to have a "large" correlation effect. (Medium effect is 0.24 – 0.36 and small effect is 0.10 – 0.23.)

Performance Measure Score Correlation

Measure	IPFs	Correlation
Follow-Up After Hospitalization 7-day (7/1/2014 – 6/30/2015)	1,145	0.34312
Follow-Up After Hospitalization 30-day (7/1/2014 – 6/30/2015)	1,145	0.43065
IPF All-Cause Unplanned Readmission Measure (Observed) (1/1/2013 – 12/31/2014)	1,184	-0.26059

All correlations are statistically significant at p-value < 0.0001.

- Face validity
  All of the 10 TEP members who were present for the face validity vote agreed that the measure score had face validity.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer analyzed the following exclusions:
  - Electroconvulsive therapy (ECT) or transcranial magnetic stimulation (TMS) during inpatient stay or follow-up period.
  - Pregnancy during inpatient stay
  - Secondary diagnosis of delirium
  - Principal diagnosis of schizophrenia with secondary diagnosis of dementia
- In the analysis, the developer compared the exclusion group, with all admissions, and the group without the excluded condition.
- The developer stated that ECT and TMS are rare procedures, and may be used as an alternative when failing pharmacotherapy.
- The developer noted that pregnancy (0.1), secondary diagnosis of delirium (2.0%), and schizophrenia with secondary diagnosis of dementia (3.2%) are rare occurrences, and all had lower filling rates for prescriptions.
- All exclusions were supported by the technical expert panel.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*

o *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

- To assess whether differences in performance across facilities are meaningful, developers constructed 95% confidence intervals for the performance rate for each IPF and compared these to the overall national average performance rate, respectively. They considered the facility rate to be statistically different from the average if the confidence intervals did not overlap with the national medication continuation rate.
- The developer used the 2013-2014 Medicare claims data for the analysis.
- 23.6% of the IPFs performed better than the national rate.
- 12.6% of the IPFs performed worse than the national rate.

Distribution of Facility Performance

Diagnosis	# IPFs	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
MDD	1,651	75.5	13.9	0.0	60.0	69.6	77.1	83.3	89.7	100.0
Schizophrenia	1,655	79.1	15.3	0.0	63.6	73.1	81.5	87.9	95.5	100.0
Bipolar disorder	1,658	78.3	14.4	0.0	63.9	72.5	80.0	86.4	93.5	100.0
Overall	1,694	78.0	11.1	0.0	66.7	73.6	79.6	84.4	88.3	100.0

Distribution of IPFs Compared to the National Medication Continuation Rate

Performance Categorization	Count IPFs	Percent IPFs
Total IPFs	1,694	100.0
Better than national rate	399	23.6
No different than national rate	572	33.8
Worse than national rate	213	12.6
Fewer than 75 discharges during the performance period	510	30.1

Question for the Committee:

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

2b6. Comparability of data sources/methods:

Not needed.

2b7. Missing Data

Developer states this is not applicable because the measure is based on claims data, which is reasonable since the measure is based on primary diagnosis and pharmacy claims.

Guidance from the Validity Algorithm

Specifications are consistent with evidence (Box 1)→potential threats assessed (Box 2) →empirical reliability testing (Box 3)→score level testing (Box 6)→correlation effect medium to high (Pearson) (Box 7) →Moderate certainty (Box 8a) →Moderate

The highest possible rating is HIGH.

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

Comments:

**clearly defined.
**Measure is limited to individuals continually enrolled in Part D. It excludes those Medicare beneficiaries in part C or who obtain medications through private insurance or the VA.
**Are all psychiatric medications covered under Medicare part D? Is the deductible affordable for all medications?
**The data does suggest there is a gap in care. To further determine the ambulatory setting that is or is not driving that gap in care, the data should be broken out by discharged to: 1. home setting, 2. group home setting, 3. homeless shelter, 4. SNF (I believe SNFs are excluded, correct?), 5. Other settings. This will help further define which population truly has a disparity and should be the target of the measure (maybe they all do, but in my experience at HFHS, that is not true).

2a2. Reliability Testing

Comments:

**Acceptable.
**Reliability was tested through medical chart review, 2 psychiatrists reviewed 150 medical records and determined PPV between medical record diagnosis and diagnosis on claims/discharge record. 92% of discharges had medications prescribed at discharge. "Few" discharges included the provision of medications at discharge. The PPV between medications prescribed and claims was 96%
**Test sample size of 75 seems to be reliable.
**reliability score was greater than 0.7 and the measure owner did determine if there was a difference between inclusion and exclusion of facilities with fewer than 75 discharges.
So, yes reliability testing is adequate.

2b1. Validity Specifications

Comments:

**Problematic
**This measure looks at filling prescriptions, but doesn't look at whether or not patients are taking them, or if they are taking them as prescribed. Just filling a prescription is one important step, but doesn't really predict outcome.
**Valid.

2b2. Validity Testing

Comments:

**Who is being measured, the health plan or hospital system?
**Found a -0.26 correlation between measures and all-cause readmissions. Found a positive correlation .34 and .43 with measure and 7 and 30 day post-discharge hospital follow-up. TEP confirmed face validity.
**Face validity testing was done and was adequate.
**I have concerns on whether a sample of 150 patient records is adequate to generalize for widespread use. Based upon that, I do not feel conclusions can be made about quality. Validity score for the 30 day post-hospitalization follow up had a large correlation effect and could be supported; not the 7-day follow up nor all cause hospital readmission rate.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias

Comments:

**Does not measure those without coverage.
**should the measures include long-acting injectables given more than 2 days before discharge?
**Exclusions are consistent with evidence. Missing data is not applicable.
**No.

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 required data elements are routinely collected for billing purposes but are not in electronic sources.
- The developer did not report any implementation challenges.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
**Acceptable.
**Missing everyone who doesn't have Medicare Part D.
**Data will be obtained from claims data. It is routinely generated.
**all claims based data, feasible to gather data elements.

Criterion 4: Usability and Use

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

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details

- The developer has submitted this measure to the Measures Under Consideration (MUC) list for use in the Inpatient Psychiatric Facility Quality Reporting Program (IPRQR).

Improvement results N/A

Unexpected findings (positive or negative) during implementation new measure – none reported.

Potential harms none reported

Vetting of the measure none reported

Feedback:

- In 2016, the Measure Application Partnership (MAP) recommended that this measure be refined and resubmitted prior to rulemaking because it recently completed field testing. MAP agreed that the testing results should demonstrate reliability and validity at the facility level in the hospital setting. MAP also discussed details in the measure specifications that need additional clarification such as (1) the definition of medication dispensation (2) how does the facility know the medication was dispensed? and (3) Medicare D is optional, how does this impact the measure?

***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?*
- o *Would inclusion of this measure in the Inpatient Psychiatric Facility Quality Reporting Program be reasonable? Are there other programs that might benefit from inclusion of this measure?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:

**Why is it a measure of inpatient psychiatric care if there a data element 30 days after discharge?
**Not yet.
**Data is not currently reported. The measure has completed field testing.
**The metric does appear to support continuity of care, however need more data to understand how large of an issue that is currently (numerically versus theoretically). I do feel inclusion of measure in psych facility quality reporting program is reasonable AFTER the various post-discharge ambulatory locations are analyzed to determine which one is driving that metric (all, some or all possible settings). This is important because if homeless or patient home is driving the lower figures versus group homes, then discharge to group homes should be an exclusion factor. Or else, certain institutions with large homeless population could be unrightfully penalized.

Criterion 5: Related and Competing Measures

Related or competing measures

None identified

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a <u>candidate</u> for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure has not been vetter by users or those being measured.

Pre-meeting public and member comments

- None received.

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

Measure Title: Medication Continuation Following Inpatient Psychiatric Discharge

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: 12/16/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (*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: Patient fills prescription, establishing medication continuation from the inpatient to the outpatient setting

☐ 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.12 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.

Effective interventions have been identified that can improve medication adherence during the transition from inpatient to outpatient care. Interventions that have been shown to increase medication compliance and prevent negative outcomes associated with nonadherence include patient education, enhanced therapeutic relationships, shared decision-making, and text-message reminders, with emphasis on multidimensional approaches (Douaihy, Kelly, & Sullivan, 2013; Haddad, Brain, & Scott, 2014; Hung, 2014; Kasckow & Zisook, 2008; Lanouette, Folsom, Sciolla, & Jeste, 2009; Mitchell, 2007; Sylvia et al., 2013). Interventions, including those described by the literature, can be implemented during steps 2 and 3 in the logic model to influence medication continuation in step 4. Because the denominator only includes patients who would require continued evidence-based pharmacotherapy and who have few barriers to access, this measure provides an indirect quality indicator of the treatment provided in steps 2 and 3.

1) Patient is admitted for inpatient psychiatric care → 2) Patient receives treatment and is stabilized →3) Patient is discharged with prescriptions for evidence-based medications and discharge treatment plan → 4) **Patient fills initial prescription, establishing medication continuation from the inpatient to the outpatient setting** → 5) Patient's symptoms are managed by pharmacotherapy → 6) Psychiatric decompensation and adverse outcomes such as emergency department visits, rehospitalization, and suicide are prevented.

*Douaihy, A. B., Kelly, T. M., & Sullivan, C. (2013). Medications for substance use disorders. *Social Work in Public Health*, *28*(3-4), 264-278. doi: 10.1080/19371918.2013.759031
*Haddad, P. M., Brain, C., & Scott, J. (2014). Nonadherence with antipsychotic medication in schizophrenia: Challenges and management strategies. *Patient Related Outcome Measures, 5*, 43-62. doi: 10.2147/PROM.S42735
*Hung, C. I. (2014). Factors predicting adherence to antidepressant treatment. *Current Opinion in Psychiatry, 27*(5), 344-349. doi: 10.1097/yco.0000000000000086

*Kasckow, J. W., & Zisook, S. (2008). Co-occurring depressive symptoms in the older patient with schizophrenia. *Drugs and Aging*, *25*(8), 631-647. doi: 10.2165/00002512-200825080-00002
*Lanouette, N. M., Folsom, D. P., Sciolla, A., & Jeste, D. V. (2009). Psychotropic medication nonadherence among United States Latinos: A comprehensive literature review. *Psychiatric Services*, *60*(2), 157-174. doi: 10.1176/appi.ps.60.2.157
*Mitchell, A. J. (2007). Understanding medication discontinuation in depression. *Psychiatric Times*, *24*(4).
*Sylvia, L. G., Hay, A., Ostacher, M. J., Miklowitz, D. J., Nierenberg, A. A., Thase, M. E., . . . Perlis, R. H. (2013). Association between therapeutic alliance, care satisfaction, and pharmacological adherence in bipolar disorder. *Journal of Clinical Psychopharmacology, 33*(3), 343-350. doi: 10.1097/JCP.0b013e3182900c6f

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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

Not applicable

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES) 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**	American Psychiatric Association. (2002). Practice guideline for the treatment of patients with bipolar disorder, second edition. Retrieved from http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/bipolar.pdf American Psychiatric Association. (2010a). Practice guideline for the treatment of patients with major depressive disorder, 3rd ed. Retrieved from http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf American Psychiatric Association. (2010b). Practice guideline for the treatment of patients with schizophrenia: 2nd ed. Retrieved from http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf US Department of Veterans Affairs, & US Department of Defense. (2016). Management of major depressive disorder (MDD). Retrieved from http://www.healthquality.va.gov/guidelines/MH/mdd/VADoDMDDCPGFINAL82916.pdf US Department of Veterans Affairs & US Department of Defense. (2010) VA/DOD clinical practice guideline for management of bipolar disorder in adults. Retrieved from http://www.healthquality.va.gov/guidelines/MH/bd/bd_305_full.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.	**Major Depressive Disorder** APA 2010a Guidelines *Acute Phase* "An antidepressant medication is recommended as an initial treatment choice for patients with mild to moderate major depressive disorder [I] and definitely should be provided for those with severe major depressive disorder unless ECT is planned [I]." p.17 "For most patients, a selective serotonin reuptake inhibitor (SSRI), serotonin norepinephrine reuptake inhibitor (SNRI), mirtazapine, or bupropion is optimal [I]." p.17 *Maintenance Treatment* "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]." p.19 VA/DOD 2016 Guidelines "In patients with MDD who achieve remission with antidepressant medication, treatment should be continued at the same dose for at least 6 months to decrease the risk of relapse. [A]" p.106 **Schizophrenia** APA 2010b Guidelines *Acute Phase Treatment* "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 [I]…The selection of an antipsychotic medication is frequently guided by the patient's previous experience with antipsychotics, including the degree of symptom response, past experience of side effects, and preferred route of medication administration. In choosing among these medications, the

	psychiatrist may consider the patient's past responses to treatment, the medication's side effect profile (including subjective responses, such as a dysphoric response to a medication), the patient's preferences for a particular medication based on past experience, the intended route of administration, the presence of co-morbid medical conditions, and potential interactions with other prescribed medications [I]. Finally, 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 [II]." p.11 *Stabilization Phase* "If the patient has improved with a particular medication regimen, continuation of that regimen and monitoring are recommended for at least 6 months [I]. Premature lowering of dose or discontinuation of medication during this phase may lead to a recurrence of symptoms and possible relapse." p.12 **Bipolar Disorder** APA 2002 Guidelines *Acute Phase* "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 [I]. For less ill patients, monotherapy with lithium, valproate, or an antipsychotic such as olanzapine may be sufficient [I]. Short-term adjunctive treatment with a benzodiazepine may also be helpful [II]. For mixed episodes, valproate may be preferred over lithium [II]. Atypical antipsychotics are preferred over typical antipsychotics because of their more benign side effect profile [I], with most of the evidence supporting the use of olanzapine or risperidone [II]. Alternatives include carbamazepine or oxcarbazepine in lieu of lithium or valproate [II]. Antidepressants should be tapered and discontinued if possible [I]. If psychosocial therapy approaches are used, they should be combined with pharmacotherapy [I]." p.9 "Manic or mixed episodes with psychotic features usually require treatment with an antipsychotic medication [II]." p.10 *Maintenance Treatment* "Maintenance regimens of medication are recommended following a manic episode [I]. Although few studies involving patients with bipolar II disorder have been conducted, consideration of maintenance treatment for this form of the illness is also strongly warranted [II]. The medications with the best empirical evidence to support their use in maintenance treatment include lithium [I] and valproate [I]; possible alternatives include lamotrigine [II] or carbamazepine or oxcarbazepine [II]. If one of these medications was used to achieve remission from the most recent depressive or manic episode, it generally should be continued [I]." p.11 VA/DOD 2010 Guidelines "Patients with severe mania should be treated with a combination of antipsychotics and lithium or valproate. These antipsychotics include olanzapine, quetiapine, aripiprazole, or risperidone [B] and may include and ziprasidone [I]." p.8 "Patients with severe mixed episode should be treated with a combination of antipsychotics and lithium or valproate. These antipsychotics include aripiprazole,

	olanzapine, risperidone, or haloperidol [B] and may include quetiapine or ziprasidone [I].” p.9 “Clozapine, with its more serious side effect profile, may be added to existing medications for severe mania or mixed episode if it has been successful in the past or if other antipsychotics have failed [I].” p.9 “Quetiapine, [A], lamotrigine [B], or lithium [B] monotherapy should be considered as first-line treatment for adult patients with BD depression.” p.26 *Maintenance Phase* “Patients who have had an acute manic episode should be treated for at least 6 months after the initial episode is controlled and encouraged to continue on life-long prophylactic treatment with medication. [A]” p.35
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The guideline authors did not grade the evidence or separate the grade for the evidence from the grade from the recommendation.
Provide all other grades and definitions from the evidence grading system	Not applicable
Grade assigned to the recommendation with definition of the grade	**Major Depressive Disorder** Guidelines from the American Psychiatric Association (APA; 2010a) to initiate and continue the medications in the numerator of this measure following an acute episode of MDD were graded as: I: Recommended with substantial clinical confidence Guidelines from the Department of Veterans Affairs/Department of Defense (VA/DoD; 2016) to continue the medications in the numerator of this measure for at least six months following an acute episode of MDD were graded as: A: Good evidence was found that the intervention improves important health outcomes and concludes that benefits substantially outweigh harm. B: At least fair evidence was found that the intervention improves health outcomes and concludes that benefits outweigh harm. **Schizophrenia** Guidelines from the APA (2010b) to initiate and continue the medications in the numerator of this measure following an acute episode of schizophrenia were graded as: I: Recommended with substantial clinical confidence The guideline from the APA (2010b) to use long-acting injectables for patients hospitalized for schizophrenia was graded as follows: II: Recommended with moderate clinical confidence

	Bipolar Disorder Guidelines from the APA (2002) on the various treatment approaches related to initiating and continuing the medications in the numerator of this measure following an acute episode of bipolar disorder were graded as: I: Recommended with substantial clinical confidence II: Recommended with moderate clinical confidence Guidelines from the VA/DoD (2010) on the various treatment approaches following an acute episode of bipolar disorder were graded as: B: At least fair evidence was found that the intervention improves health outcomes and concludes that benefits outweigh harm. I: Evidence that the intervention is effective is lacking, or poor quality, or conflicting, and the balance of benefits and harms cannot be determined. Guidelines from the VA/DoD (2010) on the continuation of medications in the numerator of this measure were graded as: A: Good evidence was found that the intervention improves important health outcomes and concludes that benefits substantially outweigh harm.
Provide all other grades and definitions from the recommendation grading system	APA grade I: Recommended with substantial clinical confidence APA grade II: Recommended with moderate clinical confidence APA grade III: May be recommended on the basis of individual circumstances VA/DoD grade A: Good evidence was found that the intervention improves important health outcomes and concludes that benefits substantially outweigh harm. VA/DoD grade B: At least fair evidence was found that the intervention improves health outcomes and concludes that benefits outweigh harm. VA/DoD grade I: Evidence that the intervention is effective is lacking, or poor quality, or conflicting, and the balance of benefits and harms cannot be determined.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The guidelines are evidence-based rather than expert opinion. Information regarding the quantity, quality, and consistency of the information on the treatment of MDD, bipolar disorder, and schizophrenia is based on extensive literature searches reviewed by expert workgroups and panels, which included practicing clinicians and research experts. For the treatment of MDD, the current APA guidelines were built upon literature reviews from pervious guidelines with the objective of emphasizing newer treatments. The literature search was conducted on studies published from January 1999 to December 2006. A total of 1,170 citations are reported in the current guideline (APA, 2010a). In a similar manner, the VA/DoD searched literature published from July 2000 to the end of 2006. A total of 253 citations are included in the current guideline (VA/DoD, 2010). The APA clinical guidelines for the treatment of schizophrenia were developed from a literature search conducted for the years 1994 to 2002. A total of 1,391 citations were included in the current guideline (APA, 2010b). The current APA clinical guidelines for the treatment of bipolar disorder were built upon a literature search of articles from 1992 to 2001. A total of 472 citations are included in the current guideline (APA, 2002). The VA/DoD clinical guidelines relied heavily on the APA guidelines and include 276 citations (VA/DoD, 2010).

Estimates of benefit and consistency across studies	**Major Depressive Disorder** Overall, the literature cited by the guidelines consistently found that pharmacotherapy is effective for the treatment of MDD. Several pharmacotherapies were reviewed through multiple meta-analyses (Anderson, 2000; Cipriani et al., 2005; Cipriani et al., 2009; Edwards & Anderson, 1999; Gartlehner, 2008), systematic reviews (Murdoch & Keam, 2005; Panzer, 2005), and numerous randomized trials that evaluated the efficacy and tolerability of pharmacological treatments for depression. Overall, the results of these studies indicate that SSRIs and SNRIs have relatively similar efficacies and tolerability. There is some evidence that tricyclic antidepressants (TCAs) may be more efficient for inpatient populations. SNRIs have been shown to be superior to placebo in multiple placebo-controlled studies (DeMartinis, Yeung, Entsuah, & Manley, 2007; Nemeroff, Entsuah, Benattia, Demitrack, Sloan, & Thase, 2008; Papakostas, Thase, Fava, Nelson, & Shelton, 2007; Papakostas, Homberger, & Fava, 2008; Septien-Velez, Pitrosky, Padmanabhan, Germain, & Tourian, 2007; Thase, Prtichette, Ossanna, Swindle, Xu, & Detke, 2007; Papakostas, Thase, Fava, Nelson, & Shelt, 2007). Several meta-analyses of controlled trials have documented small (4% – 10%) differences in treatment response for SNRIs compared to SSRIs (Cipriani, Barbui, Brambilla, Furukawa, Hotoph, & Geddes, 2006; Nemeroff et al., 2008; Papakostas et al., 2008; Smith, 2002; Thase, 2001; Thase et al., 2007). Alternative depression medications have been efficacious in reducing depressive symptoms compared to placebo, including bupropion (Fava, Rush, Thase, Clayton, Stahl, Pradko, & Johnston, 2005) and mirtazapine (Claghorn & Lesem, 1995; Holm & Markham, 1999). Monoamine oxidase inhibitors (MAOIs) have similar efficacy to TCAs (Clayton, McGarvey, Abouesh, & Pinkerton, 2001; Himmelhock, Thase, Mallinger, & Houck, 1991; Masand, Ashton, Gupta, & Frank, 2001; McGrath, Stewart, Harrison, Wager, & Quitkin, 1986; White, Razani, Cadow, Gelfand, Palmer, Simpson, & Sloan, 1984), particularly for patients who have not responded to other antidepressant medication (Himmelhoch, Fuchs, & Symons, 1982; Himmelhoch et al., 1991; White et al., 1984). All of the classes of pharmacotherapies evaluated in these studies are included in the numerator definition of this measure to allow for flexibility in prescribing an evidence-based treatment for MDD. **Schizophrenia** Overall, the literature cited by the guidelines consistently found that pharmacotherapy is effective for the treatment of schizophrenia. According to the APA guidelines for the treatment of schizophrenia (APA, 2010b), evidence supporting the use of typical (i.e., first-generation) antipsychotics was first established in the 1960s (Laskey, Klett, Caffey, Bennett, Rosenblum, & Hollister, 1962) and repeatedly confirmed by subsequent clinical trials (Davis, Barter, & Kane, 1989). These studies compared the efficacy of one or more antipsychotic medications to that of a sedative or a placebo, and nearly all confirmed the antipsychotic medication to be a superior treatment (APA, 2010b). Research on typical antipsychotics has decreased substantially since the development of atypical (i.e., second-generation) antipsychotics. There are a number of atypical antipsychotics that are effective in the treatment of schizophrenia. At the time of the development of the clinical guidelines, clozapine was considered a superior treatment compared to typical antipsychotics in six of eight published double-blind randomized trials (Buchanan, Brier, Kirkpatrick, Ball, & Carpenter, 1998; Essock, Hargreaves, Covell, & Goethe, 1996; Hong, Chen, Chiu, & Sim, 1997; Kane, Honigfeld, Singer, & Meltzer, 1988; Kane et al., 2001; Kumra et al., 1996; Rosenheck et al., 1997; Volavka et al., 2002). A subsequent meta-analysis of five of these studies

	confirmed that clozapine-treated patients were 2.5 times more likely to improve compared to those treated with a typical antipsychotic. Clinical trials that informed the clinical guidelines demonstrated other atypical antipsychotics to be superior to placebo and to typical antipsychotics, including risperdone (Borison, Pathiraja, Diamond & Meibach, 1992; Chouinard et al., 1993; Marder & Meibach, 1994) and olanzapine (Beasley, Sanger, Satterless, Tollefson, Tran, & Hamilton, 1996; Beasley et al., 1997; Hamilton, Revicki, Genduso, & Beasley, 1998; Lieberman et al., 2003; Tollefson et al., 1997). Quetiapine and aripiprazole were demonstrated to be superior to placebo and typical antipsychotics (Borison, Arvanitis, & Milier, 1996; Fabre, Arvanitis, Pultz, Jones, Malick & Slotnick, 1995; Marder et al., 2003; Small, Kirsch, Arvanitis, Miller, & Link, 1997), although their effectiveness at reducing negative symptoms of schizophrenia is variable (Borison et al., 1996; Fabre et al., 1995; Small et al., 1997; Marder et al., 2003). Meta-analyses of these studies suggest that the efficacy of quetiapine is similar to that of typical antipsychotics (Geddes, Freemantle, Harrison, & Bebbington, 2000; Leucht, Pitschel-Walz, Abraham, & Kissling, 1999; Leucht, Wahlbeck, Hamann, & Kissling, 2003). Studies of ziprasidone found that it is superior compared to placebo and typical antipsychotics (Daniel, Zimbroff, Potkin, Reeves, Harrigan, & Lakshminarayanan, 1999; Keck, Buffenstein, Ferguson, Feighner, Jaffe, Harrigan, & Morrissey, 1998), including significantly reducing the risk of relapse (Goff et al., 1998). All of the pharmacotherapies evaluated in these studies are included in the numerator definition of this measure to allow for flexibility in prescribing an evidence-based treatment for schizophrenia. **Bipolar Disorder** Overall, the literature cited by the guidelines consistently found that pharmacotherapy is effective for the treatment of bipolar disorder. Many studies have demonstrated the efficacy of mood stabilizers (including lithium, anticonvulsants, and typical and atypical antipsychotics) as a treatment for reducing the depressive symptoms and manic episodes associated with bipolar disorder. Five studies found lithium to be a superior treatment for bipolar disorder compared to placebo (Bowden, et al., 1994; Goodwin, Murphy, & Bunney, 1969; Schou, Juel-Nielson, Stroomgreen, & Voldky, 1954; Maggs, 1963; Strokes, Shamoian, Stoll, & Patton, 1971). It should be noted the interpretation of these results is limited due to the use of a cross-over design in four of the trials (Goodwin et al., 1969; Schou et al., 1954; Maggs, 1963; Strokes et al., 1971), non-random assignment (Goodwin et al., 1969; Strokes et al., 1971), and variability in diagnostic criteria. In trials comparing lithium to other active pharmacological agents, lithium displayed similar efficacy to carbamazepine (Lerer, Moore, Meyendorff, Cho, & Gershon, 1987; Small et al., 1991), risperidone (Segal, Berk, & Brook, 1998), olanzapine (Berk, Ichim, & Brook, 1999), chlorpromazine, and other typical antipsychotics (Johnson, Gershon, Burdock, Floyd, & Hekimian, 1971; Platman, 1970; Prien, Caffey, & Klett, 1972; Shopsin, Gershon, Thompson, & Collins, 1975; Spring, Schweid, Gray, Steinberg, & Horwitz, 1970; Takahashi, Sakuma, Itoh, K., Itoh, H., & Kurihara, 1975). Open studies (Himmelhoch & Garfinkel, 1986; Kramlinger & Post, 1989; Prien, Himmelhoch, & Kuper, 1988) and randomized active comparator-controlled studies (Bowden, 1995; Freeman, Clothier, Pazzaglia, Lesem, & Swann, 1992; Swann et al., 1997) demonstrate that lithium is an effective treatment for manic states but is less effective in the treatment of mixed states. The efficacy of anticonvulsants (e.g., divalproex, valproate, valproic acid) compared to placebo has been demonstrated in four randomized controlled trials (Bowden, et al., 1994; Brennan, Sandyk, & Borsook, 1984; Emrich, Zerssen, Kissling, Miller, & Windorder, 1981; Pope, McElroy, Keck, & Hudson, 1991) with response rates ranging from 48% to 58%.

	One randomized, placebo-controlled study has evaluated antipsychotics for the treatment of bipolar disorder. The results indicated that chlorpromazine was superior to placebo in the overall improvement of manic symptoms (Klein, 1967). Typical antipsychotics are comparable to lithium in effectiveness (Platman, 1970; Prien, et al., 1972; Shopsin, et al., 1975; Spring et al., 1970; Takahashi, 1975). Atypical antipsychotics (i.e., risperidone and ziprasidone) have been shown to be superior to placebo and similar to haloperidol in effectiveness (Sachs, 2001). All of the pharmacotherapies evaluated in these studies are included in the numerator definition of this measure to allow for flexibility in prescribing an evidence-based treatment for bipolar disorder. *American Psychiatric Association. (2002). Practice guideline for the treatment of patients with bipolar disorder, second edition. Retrieved from http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/bipolar.pdf *American Psychiatric Association. (2010a). Practice guideline for the treatment of patients with major depressive disorder, 3rd ed. Retrieved from http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf *American Psychiatric Association. (2010b). Practice guideline for the treatment of patients with schizophrenia: 2nd ed. Retrieved from http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf *Anderson, I. M. (2000). Selective serotonin reuptake inhibitors versus tricyclic antidepressants: A meta-analysis of efficacy and tolerability. *Journal of Affective Disorders, 58(*1), 19–36. doi:10.1016/s0165-0327(99)00092-0 *Beasley, C. M., Sanger, T., Satterlee, W., Tollefson, G., Tran, P., & Hamilton, S. (1996). Olanzapine versus placebo: Results of a double-blind, fixed-dose olanzapine trial. *Psychopharmacology, 124*(1-2), 159-167. doi:10.1007/bf02245617 *Beasley, C. M., Hamilton, S. H., Crawford, A. M., Dellva, M. A., Tollefson, G. D., Tran, P. V., … Beuzen, J.-N. (1997). Olanzapine versus haloperidol: Acute phase results of the international double-blind olanzapine trial. *European Neuropsychopharmacology, 7*(2), 125-137. doi:10.1016/s0924-977x(96)00392-6 *Berk, M., Ichim, L., & Brook, S. (1999). Olanzapine compared to lithium in mania: A double-blind randomized controlled trial. *International Clinical Psychopharmacology, 14*(6), 339-343. doi:10.1097/00004850-199911000-00003 *Bowden, C. L., Brugger, A. M., Swann, A. C., Calabrese, J. R., Janicak, P. G., Petty, F….Small, J. G. (1994). Efficacy of divalproex vs lithium and placebo in the treatment of mania: The Depakote Mania Study Group. *JAMA: The Journal of the American Medical Association, 271*(12), 918-924. doi:10.1001/jama.271.12.918 *Bowden, C. L. (1995). Predictors of response to divalproex and lithium. *Journal of Clinical Psychiatry, 56*(3), 25-30. *Borison, R. L., Pathiraja, A. P., Diamond, B. I., & Meibach, R. C. (1992). Risperidone: Clinical safety and efficacy in schizophrenia. *Psychopharmacological Bulletin, 28*, 213-218. *Borison, R. L., Arvanitis, L. A., & Milier, B. G. (1996). ICI 204,636, an atypical antipsychotic. *Journal of Clinical Psychopharmacology, 16*(2), 158-169. doi:10.1097/00004714-199604000-00008

*Bowden, C. L., Brugger, A. M., Swann, A. C., Calabrese, J. R., Janicak, P. G., Petty, F., Dilsaver, S. C.…Small, J. G. (1994). Efficacy of divalproex vs lithium and placebo in the treatment of mania. The Depakote Mania Study Group. *JAMA: The Journal of the American Medical Association, 271*(12), 918-924. doi:10.1001/jama.271.12.918

*Brennan, M. J. W., Sandyk, R., & Borsook, D. (1984). Use of sodium valproate in the management of affective disorders: Basic and clinical aspects. In Emrich, H. M., Okuma, T., & Muller, A. A. (Eds.), *Anticonvulsants in Affective Disorders* (pp. 56-65). Amsterdam: Excerpta Medica.

*Buchanan, R. W., Breier, A., Kirkpatrick, B., Ball, P., & Carpenter Jr., W. T. (1998). Positive and negative symptom response to clozapine in schizophrenic patients with and without the deficit syndrome. *American Journal of Psychiatry, 155*, 751-760.

*Cipriani, A., Brambilla, P., Furukawa, T. A., Geddes, J., Gregis, M., Hotopf, M., … Barbui, C. (2005). Fluoxetine versus other types of pharmacotherapy for depression. *Reviews*. doi:10.1002/14651858.cd004185.pub2

*Cipriani, A., Barbui, C., Brambilla, P., Furukawa, T. A., Hotopf, M., & Geddes, J. R. (2006). Are all antidepressants really the same? The case of Fluoxetine. *The Journal of Clinical Psychiatry, 67*(06), 850–864. doi:10.4088/jcp.v67n0601

*Cipriani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J. P., Churchill, R., … Barbui, C. (2009). Comparative efficacy and acceptability of 12 new-generation antidepressants: A multiple-treatments meta-analysis. *The Lancet, 373*(9665), 746–758. doi:10.1016/s0140-6736(09)60046-5

*Chouinard, G., Jones, B., Remington, G., Bloom, D., Addington, D., MacEwan, G. W., … Arnott, W. (1993). A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients. *Journal of Clinical Psychopharmacology, 13*(1), 25???40. doi:10.1097/00004714-199302000-00004

*Claghorn, J. L., & Lesem, M. D. (1995). A double-blind placebo-controlled study of Org 3770 in depressed outpatients. *Journal of Affective Disorders, 34*(3), 165-171. doi:10.1016/0165-0327(95)00014-e

*Clayton, A. H., McGarvey, E. L., Abouesh, A. I., & Pinkerton, R. C. (2001). Substitution of an SSRI with bupropion sustained release following SSRI-induced sexual dysfunction. *The Journal of Clinical Psychiatry, 62*(3), 185-190. doi:10.4088/jcp.v62n0309

*Daniel, D. G., Zimbroff, D. L., Potkin, S. G., Reeves, K. R., Harrigan, E. P., Lakshminarayanan, M. (1999). Ziprasidone 80 mg/day and 160 mg/day in the acute exacerbation of schizophrenia and schizoaffective disorder: A 6-week placebo-controlled trial. *Neuropsychopharmacology, 20(*5), 491-505. doi:10.1016/s0893-133x(98)00090-6

*Davis, J. M., Barter, J. T., Kane, J. M. (1989). Antipsychotic drugs. In H. I. Kaplan & B. J. Sadock (Eds.), *Comprehensive Textbook of Psychiatry, 5th ed* (pp. 1591-1626). Baltimore, MD. Williams & Wilkins.

*DeMartinis, N. A., Yeung, P. P., Entsuah, R., & Manley, A. L. (2007). A double-blind, placebo-controlled study of the efficacy and safety of desvenlafaxine succinate in the treatment of major depressive disorder. *The Journal of Clinical Psychiatry, 68*(05), 677-688. doi:10.4088/jcp.v68n0504

*Edwards, J. G., & Anderson, I. (1999). Systematic review and guide to selection of selective serotonin reuptake inhibitors. *Drugs, 57*(4), 507-533. doi:10.2165/00003495-199957040-00005

*Emrich, H. M., Zerssen, D., Kissling, W., Miller, H.-J., & Windorfer, A. (1980). Effect of sodium valproate on mania. *Archive for Psychiatrie and Nervenkrankheiten, 229*(1), 1-16. doi:10.1007/bf00343800

*Emrich, H. M., Zihl, J., Raptis, C., & Wendl, A. (1990). Reduced dark-adaptation: An indication of lithium's neuronal action in humans. *American Journal of Psychiatry, 147*(5), 629-631. doi: 10.1176/ajp.147.5.629

*Essock, S. M., Hargreaves, W. A. Covell, N. H., & Goethe, J. (1996). Clozapine's effectiveness for patients in state hospitals: Results from a randomized trial. *Psychopharmacological Bulletin, 32*, 683-697.

*Fabre, L. F., Arvanitis, L., Pultz, J., Jones, V. M., Malick, J. B., & Slotnick, V. B. (1995). ICI 204,636, a novel, atypical antipsychotic: Early indication of safety and efficacy in patients with chronic and subchronic schizophrenia. *Clinical Therapeutics, 17*(3), 366-378. doi:10.1016/0149-2918(95)80102-2

*Fava, M., Rush, A. J., Thase, M. E., Clayton, A., Stahl, S. M., Pradko, J. F., & Johnston, J. A. (2005). 15 years of clinical experience with bupropion HCl. *The Primary Care Companion to The Journal of Clinical Psychiatry, 07*(03), 106-113. doi:10.4088/pcc.v07n0305

*Freeman, T. W., Clothier, J. L., Pazzaglia, P., Lesem, M. D., & Swann, A. C. (1992). A double-blind comparison of valproate and lithium in the treatment of acute mania. *American Journal of Psychiatry, 149*(1), 108–111. doi:10.1176/ajp.149.1.108

*Gartlehner, G. (2008). Comparative benefits and harms of second-generation antidepressants: Background paper for the American College of Physicians. *Ann Intern Med, 149*(10), 734. doi:10.7326/0003-4819-149-10-200811180-00008

*Geddes, J., Freemantle, N., Harrison, P., & Bebbington, P. (2000). Atypical antipsychotics in the treatment of schizophrenia: Systematic overview and meta-regression analysis. *British Medical Journal, 321*(7273), 1371-1376. doi:10.1136/bmj.321.7273.1371

*Goff, D. C., Posever, T., Herz, L., Simmons, J., Kletti, N., Lapierre, K., … Ko, G. N. (1998). An exploratory haloperidol-controlled dose-finding study of ziprasidone in hospitalized patients with schizophrenia or schizoaffective disorder. *Journal of Clinical Psychopharmacology, 18*(4), 296-304. doi:10.1097/00004714-199808000-00009

*Goodwin, F. K. (1969). Lithium-carbonate treatment in depression and mania. *Archives of General Psychiatry, 21*(4), 486. doi:10.1001/archpsyc.1969.01740220102012

*Hamilton, S. H., Revicki, D. A., Genduso, L. A., Beasley, C. M. (1998). Olanzapine versus placebo and haloperidol: Quality of life and efficacy results of the North American Double-blind Trial. *Neuropsychopharmacology, 18*(1), 41-49. doi:10.1016/s0893-133x(97)00111-5

*Himmelhoch, J. M., Fuchs, C. Z., & Symons, B. J. (1982). A double-blind study of tranylcypromine treatment of major anergic depression. *The Journal of Nervous and Mental Disease, 170*(10), 628-634. doi:10.1097/00005053-198210000-00007

*Himmelhoch, J. M. & Garfinkel, M. E. (1986). Sources of lithium resistance in mixed mania. *Psychopharmacological Bulletin, 22*, 613-620.

*Himmelhoch, J. M., Thase, M. E., Mallinger, A. G., & Houck, P. (1991). Tranylcypromine versus imipramine in anergic bipolar depression. *American Journal of Psychiatry, 148*(7), 910-916. doi:10.1176/ajp.148.7.910

*Holm, K. J., & Markham, A. (1999). Mirtazapine. *Drugs, 57*(4), 607–631. doi:10.2165/00003495-199957040-00010

*Hong, C. J., Chen, J. Y., Chiu, H. J., & Sim, C. B. (1997). A double-blind comparative study of clozapine versus chlorpromazine on Chinese patients with treatment-refractory schizophrenia. *International Clinical Psychopharmacology, 12*(3), 123-130. doi:10.1097/00004850-199705000-00001

*Johnson, G., Gershon, S., Burdock, E. I., Floyd, A., & Hekimian, L. (1971). Comparative effects of lithium and chlorpromazine in the treatment of acute manic states. *The British Journal of Psychiatry, 119*(550), 267–276. doi:10.1192/bjp.119.550.267

*Kane, J., Honigfeld, G., Singer, J., & Meltzer, H. (1988). Clozapine for the treatment-resistant schizophrenic. *Archives of General Psychiatry, 45*(9), 789. doi:10.1001/archpsyc.1988.01800330013001

*Kane, J. M., Marder, S. R., Schooler, N. R., Wirshing, W. C., Umbricht, D., Baker, R. W., … Borenstein, M. (2001). Clozapine and haloperidol in moderately refractory schizophrenia. *Archives of General Psychiatry, 58*(10), 965. doi:10.1001/archpsyc.58.10.965

*Keck Jr, P., Buffenstein, A., Ferguson, J., Feighner, J., Jaffe, W., Harrigan, E. P., & Morrissey, M. R. (1998). Ziprasidone 40 and 120 mg/day in the acute exacerbation of schizophrenia and schizoaffective disorder: A 4-week placebo-controlled trial. *Psychopharmacology, 140*(2), 173–184. doi:10.1007/s002130050755

*Klein, D. F. (1967). Importance of psychiatric diagnosis in prediction of clinical drug effects. *Archives of General Psychiatry, 16*(1), 118. doi:10.1001/archpsyc.1967.01730190120016

*Kramlinger, K. G., & Post, R. M. (1989). Adding lithium carbonate to carbamazepine: Antimanic efficacy in treatment-resistant mania. *Acta Psychiatrica Scandinavica, 79*(4), 378–385. doi:10.1111/j.1600-0447.1989.tb10273.x

*Kumra, S., Frazier, J. A., Jacobsen, L. K., McKenna, K., Gordon, C. T., Lenane, M. C.,…Rapoport, J. L. (1996). Childhood-onset schizophrenia. *Archives of General Psychiatry, 53*(12), 1090. doi:10.1001/archpsyc.1996.01830120020005

*Laskey, J. J., Klett, C. J., Caffey, E. M. Jr., Bennett, J. L., Rosenblum, M. P., and Hollister, L. E. (1962). Drug treatment of schizophrenic patients: A comprehensive evaluation of chlorpromazine, chlorprothixene, fluphenazine, reserpine, thioridazine, and trifluprom azine. *Diseases of the Nervous System,* 23. 698-706.

*Lerer, B., Moore, N., Meyendorff, E., Cho, S. R., & Gershon, S. (1987). Carbamazepine versus lithium in mania: A double-blind study. *Journal of Clinical Psychiatry, 48,* 89-93.

*Leucht, S., Pitschel-Walz, G., Abraham, D., & Kissling, W. (1999). Efficacy and extrapyramidal side-effects of the new antipsychotics olanzapine, quetiapine, risperidone, and sertindole compared to conventional antipsychotics and placebo. A meta-analysis of randomized controlled trials. *Schizophrenia Research, 35*(1), 51-68. doi:10.1016/s0920-9964(98)00105-4

*Leucht, S., Wahlbeck, K., Hamann, J., & Kissling, W. (2003). New generation antipsychotics versus low-potency conventional antipsychotics: A systematic review and meta-analysis. *The Lancet, 361*(9369), 1581–1589. doi:10.1016/s0140-6736(03)13306-5

*Maggs, R. (1963). Treatment of manic illness with lithium carbonate. *The British Journal of Psychiatry, 109*(458), 56–65. doi:10.1192/bjp.109.458.56

*Marder, S. R. & Meibach, R. C. (1994). Risperidone in the treatment of schizophrenia. *American Journal of Psychiatry, 151,* 825-835.

*Marder, S. R., McQuade, R. D., Stock, E., Kaplita, S., Marcus, R., Safferman, A. Z., … Iwamoto, T. (2003). Aripiprazole in the treatment of schizophrenia: Safety and

tolerability in short-term, placebo-controlled trials. *Schizophrenia Research, 61*(2-3), 123–136. doi:10.1016/s0920-9964(03)00050-1

*Masand, P. S., Ashton, A. K., Gupta, S., & Frank, B. (2001). Sustained-release bupropion for selective serotonin reuptake inhibitor-induced sexual dysfunction: A randomized, double-blind, placebo-controlled, parallel-group study. *American Journal of Psychiatry, 158*(5), 805–807. doi:10.1176/appi.ajp.158.5.805

*McGrath, P. J., Stewart, J. W., Harrison, W., Wager, S., & Quitkin, F. M. (1986). Phenelzine treatment of melancholia. *Journal of Clinical Psychiatry, 47*, 420-422.

*Muller, A. A. & Stoll, K. D. (1984). Anticonvulsants in affective disorders. In Emrich, H. M., Okuma, T., & Muller, A. A. (Eds.) *Carbamazepine and Oxcarbazepine in the treatment of manic syndromes: Studies in Germany* (pp. 134-147). Amsterdam, The Netherlands: Exerpta Medica.

*Murdoch, D. & Keam, S. J. (2005). Escitalopram: A review of its use in the management of major depressive disorder. *Drugs, 65*,2379-2404. doi: 10.2165/00003495-200565160-00013

*Nemeroff, C. B., Entsuah, R., Benattia, I., Demitrack, M., Sloan, D. M., & Thase, M. E. (2008). Comprehensive analysis of remission (COMPARE) with Venlafaxine versus SSRIs. *Biological Psychiatry, 63*(4), 424-434. doi:10.1016/j.biopsych.2007.06.027

*Panzer, M. J. (2005). Are SSRIs really more effective for anxious depression? *Annals of Clinical Psychiatry, 17*(1), 23-29. doi:10.1080/10401230590905317

*Papakostas, G. I., Thase, M. E., Fava, M., Nelson, J. C., & Shelton, R. C. (2007). Are antidepressant drugs that combine serotonergic and noradrenergic mechanisms of action more effective than the selective serotonin reuptake inhibitors in treating major depressive disorder? A meta-analysis of studies of newer agents. *Biological Psychiatry, 62*(11), 1217-1227. doi:10.1016/j.biopsych.2007.03.027

*Papakostas, G., Homberger, C., & Fava, M. (2008). A meta-analysis of clinical trials comparing mirtazapine with selective serotonin reuptake inhibitors for the treatment of major depressive disorder. *Journal of Psychopharmacology, 22*(8), 843–848. doi:10.1177/0269881107083808

*Platman, S. R. (1970). A comparison of lithium carbonate and chlorpromazine in mania. *American Journal of Psychiatry, 127*(3), 351–353. doi:10.1176/ajp.127.3.351

*Pope, Jr., H. G., McElroy, S. L., Keck, Jr., P. E., & Hudson, J. I. (1991). Valproate in the treatment of acute mania: A placebo-controlled study. *Archives of General Psychiatry, 48*, 62-68.

*Prien, R. F., Caffey Jr., E. M., Klett, C. J. (1972). Comparison of lithium carbonate and chlorpromazine in the treatment of mania: Report of the Veterans Administration and National Institute of Mental Health Collaborative Study Group. *Archives of General Psychiatry, 26*(2), 146. doi:10.1001/archpsyc.1972.01750200050011

*Prien, R. F., Himmelhoch, J. M., & Kupfer, D. J. (1988). Treatment of mixed mania. *Journal of Affective Disorders, 15*(1), 9–15. doi:10.1016/0165-0327(88)90003-1

*Rosenheck, R., Cramer, J., Xu, W., Thomas, J., Henderson, W., Frisman, L., … Charney, D. (1997). A comparison of clozapine and haloperidol in hospitalized patients with refractory schizophrenia. *New England Journal of Medicine, 337*(12), 809–815. doi:10.1056/nejm199709183371202

*Sachs, G. S. (2001). Emerging data: Atypical antipsychotics in bipolar disorder. In *Program and Abstracts of the 52nd Institute on Psychiatric Services*. Washington, D. C.: American Psychiatric Association.

*Schou, M., Juel-Nielsen, N., Stromgren, E., & Voldby, H. (1954). The treatment of manic psychoses by the administration of lithium salts. *Journal of Neurology, Neurosurgery & Psychiatry, 17*(4), 250–260. doi:10.1136/jnnp.17.4.250

*Shopsin, B., Gershon S., Thompson, H., & Collins, P. (1975). Psychoactive drugs in mania. *Archives of General Psychiatry, 32*(1), 34. doi:10.1001/archpsyc.1975.01760190036004

*Segal, J., Berk, M., & Brook, S. (1998). Risperidone compared with both lithium and haloperidol in mania: A double-blind randomized controlled trial. *Clinical Neuropharmacology, 21*, 176-180.

*Septien-Velez, L., Pitrosky, B., Padmanabhan, S. K., Germain, J.-M., & Tourian, K. A. (2007). A randomized, double-blind, placebo-controlled trial of desvenlafaxine succinate in the treatment of major depressive disorder. *International Clinical Psychopharmacology, 22*(6), 338–347. doi:10.1097/yic.0b013e3281e2c84b

*Small, J. G. (1991). Carbamazepine compared with lithium in the treatment of mania. *Archives of General Psychiatry, 48*(10), 915. doi:10.1001/archpsyc.1991.01810340047006

*Small, J. G., Hirsch, S. R., Arvanitis, L. A., Miller, B. G., & Link, C. G. (1997). Quetiapine in patients with schizophrenia. *Archives of General Psychiatry, 54*(6), 549. doi:10.1001/archpsyc.1997.01830180067009

*Smith, D. (2002). Efficacy and tolerability of venlafaxine compared with selective serotonin reuptake inhibitors and other antidepressants: A meta-analysis. *The British Journal of Psychiatry, 180*(5), 396–404. doi:10.1192/bjp.180.5.396

*Spring, G., Schweid, D., Gray, C., Steinberg, J., & Horwitz, M. (1970). A double-blind comparison of lithium and chlorpromazine in the treatment of manic states. *American Journal of Psychiatry, 126*(9), 1306–1310. doi:10.1176/ajp.126.9.1306

*Stokes, P., Shamoian, C., Stoll, P., & Patton, M. (1971). Efficacy of lithium as acute treatment of manic-depressive illness. *The Lancet, 297*(7713), 1319–1325. doi:10.1016/s0140-6736(71)91886-1

*Swann, A. C., Bowden, C. L., Morris, D., Calabrese, J. R., Petty, F.,…Davis, J. M. (1997). Depression during mania: Treatment response to lithium or divalproex. *Archives of General Psychiatry, 54*(1), 37. doi:10.1001/archpsyc.1997.01830130041008

*Takahashi, R. (1975). Comparison of efficacy of lithium carbonate and chlorpromazine in mania. *Archives of General Psychiatry, 32*(10), 1310. doi:10.1001/archpsyc.1975.01760280108010

*Thase, M. E., Pritchett, Y. L., Ossanna, M. J., Swindle, R. W., Xu, J., & Detke, M. J. (2007). Efficacy of Duloxetine and selective serotonin reuptake inhibitors. *Journal of Clinical Psychopharmacology, 27*(6), 672–676. doi:10.1097/jcp.0b013e31815a4412

*Thase, M. E. (2001). Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. *The British Journal of Psychiatry, 178*(3), 234–241. doi:10.1192/bjp.178.3.234

*Tollefson, G. D., Beasley Jr., C. M., Tran, P. V., Street, J. S., Krueger, J. A., Tamura, R. N.,… Thieme, M. E. (1997). Olanzapine versus haloperidol in the treatment of schizophrenia and schizoaffective and schizophreniform disorders: Results of an international collaborative trial. *American Journal of Psychiatry, 154*(4), 457–465. doi:10.1176/ajp.154.4.457

*Volavka, J., Czobor, P., Sheitman, B., Lindenmayer, J.-P., Citrome, L., McEvoy, J. P., … Lieberman, J. A. (2002). Clozapine, olanzapine, risperidone, and haloperidol in the

	treatment of patients with chronic schizophrenia and schizoaffective disorder. *American Journal of Psychiatry, 159*(2), 255–262. doi:10.1176/appi.ajp.159.2.255 *White, K., Razani, J., Cadow, B., Gelfand, R., Palmer, R., Simpson, G., & Sloane, R. B. (1984). Tranylcypromine vs nortriptyline vs placebo in depressed outpatients: A controlled trial. *Psychopharmacology, 82*(3), 258–262. doi:10.1007/bf00427786 *US Department of Veterans Affairs, & US Department of Defense. (2009). management of major depressive disorder (MDD). Retrieved from http://www.healthquality.va.gov/mdd/mdd_full09_c.pdf
What harms were identified?	Medications associated with the treatment of MDD, schizophrenia, and bipolar disorder have been shown to reduce negative symptoms, and the clinical guidelines indicate that the benefits outweigh harms for patients with severe mental illness. However, many of the medications require careful monitoring to avoid harmful side effects. Clinicians prescribing medications for the treatment of these disorders must consider the specific medication and the side effects that might occur. These considerations may vary given a patient's clinical and personal characteristics, as well as the expected improvement in the patient's outcomes. The implementation of this measure will provide the important benefit of quality improvement by helping to identify patients who do not continue their pharmacotherapy post-discharge. Improved medication continuation would help reduce the risk of symptom relapse, prevent future depressive/manic/psychotic episodes, decrease re-hospitalization and suicide rates, and improve the quality of care for individuals with major depressive disorder, schizophrenia, and bipolar disorder.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Since the development of the clinical guidelines for schizophrenia, the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Project compared the longer-term effects and usefulness of typical (perphenazine, fluphenazine decanoate) and atypical (olanzapine, quetiapine, risperidone, ziprasidone, clozapine) antipsychotics. A study based on data from that project found that perphenazine, a typical antipsychotic, was equally as effective as the atypical antipsychotics quetiapine, risperidone, and ziprasidone (Lieberman, et al., 2010). This finding further supports the inclusion of both types of antipsychotics in the numerator definition for schizophrenia in this measure. *Lieberman, J. A., Tollefson, G., Tohen, M., Green, A. I., Gur, R. E., Kahn, R., … Hamer, R. M. (2003). Comparative efficacy and safety of atypical and conventional antipsychotic drugs in first-episode psychosis: A randomized, double-blind trial of olanzapine versus haloperidol. *American Journal of Psychiatry, 160*(8), 1396–1404. doi:10.1176/appi.ajp.160.8.1396

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.

Not applicable

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

Not applicable

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

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

Med_Continuation_nqf_evidence_attachment.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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 PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. (Describe how and from whom their input was obtained.)
IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

The aim of the proposed measure is to address gaps in continuity of pharmaceutical treatment during the transition from inpatient care to outpatient care. Pharmacotherapy is the primary form of treatment for most patients discharged from an inpatient psychiatric facility (IPF) for major depressive disorder (MDD), schizophrenia, or bipolar disorder. The measure focuses on medication continuation because it is an essential step in medication adherence.

Medication continuation is particularly important in the psychiatric patient population because psychotropic medication discontinuation can have a range of adverse effects, from mild withdrawal to life-threatening autonomic instability and psychiatric decompensation (Ward & Schwartz, 2013). Patients with MDD who do not remain on prescribed medication are more likely to have negative health outcomes, such as relapse and readmission, decreased quality of life, and increased healthcare costs. If untreated, MDD can contribute to or worsen chronic medical disorders (Geddes et al., 2003; Glue et al., 2010). The literature shows that among patients with schizophrenia, those who were "good compliers" according to the Medication Adherence Rating Scale had better outcomes in terms of rehospitalization rates and medication maintenance (Jaeger et al., 2012). Among patients with bipolar disorder, medication adherence was significantly associated with reduction in manic symptoms (Sylvia et al., 2013), while non-adherence was associated with increased suicide risk (OR 10.8, CI 1.57–74.4; Gonzalez-Pinto et al., 2006).

Current facility-level performance indicates that there is a clear quality gap. Using 2013–2014 Medicare claims data, we found that there is about a 22 percentage point difference between the 10th and 90th percentiles (66.7%-88.3%) and a median score of

79.6%. By calculating the facility-level rates of medication continuation in Medicare FFS claims data, this measure can provide valuable information on areas where care transitions to the outpatient setting can be improved.

*Geddes, J. R., Carney, S. M., Davies, C., Furukawa, T. A., Kupfer, D. J., Frank, E., & Goodwin, G. M. (2003). Relapse prevention with antidepressant drug treatment in depressive disorders: A systematic review. The Lancet, 361(9358), 653–661. doi:10.1016/s0140-6736(03)12599-8
*Glue, P., Donovan, M. R., Kolluri, S., & Emir, B. (2010). Meta-analysis of relapse prevention antidepressant trials in depressive disorders. Australian and New Zealand Journal of Psychiatry, 44(8), 697-705. doi: 10.3109/00048671003705441
*Gonzalez-Pinto, A., Mosquera, F., Alonso, M., López, P., Ramírez, F., Vieta, E., & Baldessarini, R. J. (2006). Suicidal risk in bipolar I disorder patients and adherence to long-term lithium treatment. Bipolar Disorders, 8(5p2), 618–624. doi:10.1111/j.1399-5618.2006.00368.x
*Jaeger, S., Pfiffner, C., Weiser, P., Kilian, R., Becker, T., Langle, G., . . . Steinert, T. (2012). Adherence styles of schizophrenia patients identified by a latent class analysis of the Medication Adherence Rating Scale (MARS): A six-month follow-up study. Psychiatry Research, 200(2-3), 83-88. doi: 10.1016/j.psychres.2012.03.033
*Sylvia, L. G., Hay, A., Ostacher, M. J., Miklowitz, D. J., Nierenberg, A. A., Thase, M. E., . . . Perlis, R. H. (2013). Association between therapeutic alliance, care satisfaction, and pharmacological adherence in bipolar disorder. Journal of Clinical Psychopharmacology, 33(3), 343-350. doi: 10.1097/JCP.0b013e3182900c6f
*Ward, M., & Schwartz, A. (2013). Challenges in pharmacologic management of the hospitalized patient with psychiatric comorbidity. Journal of Hospital Medicine, 8(9), 523–529. doi:10.1002/jhm.2059

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 (4b) under Usability and Use.*
HSAG calculated the performance scores using 2013–2014 Medicare claims data. Across more than 1,600 IPFs, performance varies between high- and low-performing facilities for each of the three diagnoses within the follow-up period. Median performance scores by diagnosis are 77.1% for MDD, 81.5% for schizophrenia, and 80.0% for bipolar disorder. The overall distribution of scores is provided below.

IPFs//Mean//SD//Min//10th Pctl//Lower Quartile//Median//Upper Quartile// 90th Pctl//Max
1,694// 78.0//11.1//0.0//66.7//73.6//79.6//84.4//88.3//100

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 because performance data are presented in 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 (4b) under Usability and Use.*
In order to assess whether disparities in measure performance exist between subpopulations of the measure cohort, we used the method employed by the Agency for Healthcare Research and Quality (AHRQ) for the National Healthcare Quality and Disparities Report. Two criteria are applied to determine meaningful differences between the performance for a reference group and another population group.

1. The difference is statistically significant ($p<0.05$ on a two-tailed test).
2. Relative difference between the population and reference group is at least 10%.

Results may be interpreted as better, worse, or the same as a reference group. In the disparities analyses for the measure, male is the reference group for gender, white is the reference group for race/ethnicity, the age group 65-74 is the reference group for

age, and Medicare only is the reference group for dual enrolled status. The analyses found that black patients have significantly worse rates of medication continuation than the reference group. Dually enrolled patients have significantly better rates of medication continuation than patients enrolled in only Medicare. There are no differences in performance in any of the age groups.

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

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

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

De.5. **Subject/Topic Area** *(check all the areas that apply):*

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

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.)*
Measure-specific webpage not available at the time of endorsement submission.

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:** Med_Continuation_Data_Dictionary_161216.xlsx

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

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

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The numerator for this measure includes:
1. Discharges with a principal diagnosis of MDD in the denominator population for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge
2. Discharges with a principal diagnosis of schizophrenia in the denominator population for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge
3. Discharges with a principal diagnosis of bipolar disorder in the denominator population for which patients were dispensed evidence-based outpatient medication within 2 days prior to discharge through 30 days post-discharge

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 following are the evidence-based medications by class for the treatment of MDD, schizophrenia, and bipolar disorder. The route of administration includes all oral formulations and the long-acting (depot) injectable of the medications listed in this section, except where noted. Active ingredients for the oral medications listed are limited to oral, buccal, sublingual, and translingual formulations only. Obsolete drug products are excluded from NDCs with an inactive date more than three years prior to the beginning of the measurement period.

MEDICATIONS FOR TREATMENT OF MDD
Monoamine Oxidase Inhibitors
-isocarboxazid
-phenelzine
-selegiline (transdermal patch)
-tranylcypromine

Selective Serotonin Reuptake Inhibitors (SSRI)
-citalopram
-escitalopram
-fluoxetine
-fluvoxamine
-paroxetine
-sertraline

Serotonin Modulators
-nefazodone
-trazodone
-vilazodone
-vortioxetine

Serotonin Norepinephrine Reuptake Inhibitors (SNRI)
-desvenlafaxine
-duloxetine
-levomilnacipran
-venlafaxine

Tricyclic and Tetracyclic Antidepressants

-amitriptyline
-amoxapine
-clomipramine
-desipramine
-doxepin
-imipramine
-maprotiline
-nortriptyline
-protriptyline
-trimipramine

Other Antidepressants
-bupropion
-mirtazapine

Psychotherapeutic Combinations
-amitriptyline-chlordiazepoxide
-amitriptyline-perphenazine
-fluoxetine-olanzapine

MEDICATIONS FOR TREATMENT OF SCHIZOPHRENIA
First-generation Antipsychotics
-chlorpromazine
-fluphenazine
-haloperidol
-haloperidol lactate
-loxapine succinate
-molindone
-perphenazine
-pimozide
-prochlorperazine
-thioridazine
-thiothixene
-trifluoperazine

Second-generation (Atypical) Antipsychotics
-aripiprazole
-asenapine
-brexpiprazole
-cariprazine
-clozapine
-iloperidone
-lurasidone
-olanzapine
-paliperidone
-quetiapine
-risperidone
-ziprasidone

Psychotherapeutic Combinations
-amitriptyline-perphenazine
-fluoxetine-olanzapine

Long-Acting (Depot) Injectable Antipsychotics
-fluphenazine decanoate
-haloperidol decanoate
-aripiprazole
-aripiprazole lauroxil
-olanzapine pamoate
-paliperidone palmitate (1-month extended-release injection)
-paliperidone palmitate (3-month extended-release injection)
-risperidone microspheres

MEDICATIONS FOR TREATMENT OF BIPOLAR DISORDER
Anticonvulsants
-carbamazepine
-divalproex sodium
-lamotrigine
-valproic acid

First-generation Antipsychotics
-chlorpromazine
-fluphenazine
-haloperidol
-haloperidol lactate
-loxapine succinate
-molindone
-perphenazine
-pimozide
-prochlorperazine
-thioridazine
-thiothixene
-trifluoperazine

Second-generation (Atypical) Antipsychotics
-aripiprazole
-asenapine
-brexpiprazole
-cariprazine
-clozapine
-iloperidone
-lurasidone
-olanzapine
-paliperidone
-quetiapine
-risperidone
-ziprasidone

Lithium Salts
-litium
-lithium carbonate
-lithium citrate

Psychotherapeutic Combinations
-fluoxetine-olanzapine

Long-acting (depot) Injectable Antipsychotics
-fluphenazine decanoate
-haloperidol decanoate
-aripiprazole
-aripiprazole lauroxil
-olanzapine pamoate
-paliperidone palmitate (1-month extended-release injection)
-paliperidone palmitate (3-month extended-release injection)
-risperidone microspheres

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The target population for this measure is Medicare fee-for-service (FFS) beneficiaries with Part D coverage aged 18 years and older discharged from an inpatient psychiatric facility with a principal diagnosis of MDD, schizophrenia, or bipolar disorder.

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 for this measure includes patients discharged from an IPF:
1. With a principal diagnosis of MDD, schizophrenia, or bipolar disorder (ICD codes provided below).
2. 18 years of age or older at admission.
3. Enrolled in Medicare fee-for-service Part A and Part B during the index admission and Parts A, B, and D at least 30-days post-discharge.
4. Alive at discharge and alive during the follow-up period.
5. With a discharge status code indicating that they were discharged to home or home health care.

ICD-9-CM and ICD-10-CM codes to identify MDD, schizophrenia, and bipolar disorder:
MDD
ICD-9-CM:
296.20, 296.21, 296.22, 296.23, 296.24, 296.25,
296.30, 296.31, 296.32, 296.33, 296.34, 296.35,
298.0, 311

ICD-10-CM:
F32.0, F32.1, F32.2, F32.3, F32.4, F32.9, F33.0,
F33.1, F33.2, F33.3, F33.40, F33.41, F33.9

Schizophrenia
ICD-9-CM:
295, 295.0, 295.00, 295.01, 295.02, 295.03, 295.04, 295.05,
295.1, 295.10, 295.11, 295.12, 295.13, 295.14, 295.15,
295.2, 295.20, 295.21, 295.22, 295.23, 295.24, 295.25,
295.3, 295.30, 295.31, 295.32, 295.33, 295.34, 295.35,
295.4, 295.40, 295.41, 295.42, 295.43, 295.44, 295.45,
295.5, 295.50, 295.51, 295.52, 295.53, 295.54, 295.55,
295.6, 295.60, 295.61, 295.62, 295.63, 295.64, 295.65,
295.7, 295.70, 295.71, 295.72, 295.73, 295.74, 295.75,
295.8, 295.80, 295.81, 295.82, 295.83, 295.84, 295.85,
295.9, 295.90, 295.91, 295.92, 295.93, 295.94, 295.95

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

Bipolar disorder
ICD-9-CM:
296.00, 296.01, 296.02, 296.03, 296.04, 296.05, 296.06,
296.10, 296.11, 296.12, 296.13, 296.14, 296.15, 296.16,
296.40, 296.41, 296.42, 296.43, 296.44, 296.45, 296.46,
296.50, 296.51, 296.52, 296.53, 296.54, 296.55, 296.56,
296.60, 296.61, 296.62, 296.63, 296.64, 296.65, 296.66,
296.7, 296.80, 296.81, 296.82, 296.89

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, F31.81,
F31.89, F31.9, F32.8

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
The denominator for this measure excludes discharged patients who:
1. Received Electroconvulsive Therapy (ECT) during the inpatient stay or follow-up period.
2. Received Transcranial Magnetic Stimulation (TMS) during the inpatient stay or follow-up period.
3. Were pregnant during the inpatient stay.
4. Had a secondary diagnosis of delirium.
5. Had a principal diagnosis of schizophrenia with a secondary diagnosis of dementia.

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.)*
1. ECT During Inpatient Stay or Follow-Up Period
Rationale: Some patients who receive ECT during the inpatient stay or follow-up period may have failed pharmacotherapy and would not fill an evidence-based prescription post-discharge.
Source: Identified from Part A and Part B claims data if treatment occurred on a date between the admission date and 30 days post-discharge.
2. TMS During Inpatient Stay or Follow-Up Period
Rationale: Some patients who receive TMS during the inpatient stay or follow-up period may have failed pharmacotherapy and would not fill an evidence-based prescription post-discharge.
Source: Identified from Part A and Part B claims data if treatment occurred on a date between the admission date and 30 days post-discharge.
3. Pregnant During Inpatient Stay
Rationale: Some of the evidence-based medications for the treatment of MDD, schizophrenia, and bipolar disorder are contraindicated during pregnancy.
Source: Identified from Part A claims data from the index admission.
4. Secondary Diagnosis of Delirium
Rationale: Some of the evidence-based medications for the treatment of MDD, schizophrenia, and bipolar disorder are contraindicated for patients with delirium.
Source: Identified from Part A claims data from the index admission.
5. Principal Diagnosis of Schizophrenia with Secondary Diagnosis of Dementia
Rationale: APA Practice guidelines suggest caution in the use of antipsychotics in dementia patients so not all dementia patients would fill an evidence-based medication (antipsychotic) following discharge for schizophrenia.
Source: Identified from Part A claims data from the index admission.

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 not stratified.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**
Rate/proportion
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
Denominator:
1. Pull all IPF discharges from the Part A data.
2. Include IPF discharges for patients who were at least 18 years of age at admission.
3. Identify interim claims having the same beneficiary, provider, admission dates or having an admission date within 1 day of the discharge date of the previous claim, and having a discharge status code of "Still patient." Collapse or combine the interim claims into one hospital stay using the admission date from the earliest claim and the discharge date from the latest claim. The data values from the latest claim are used for the newly combined hospital stay.
4. De-duplicate the IPF inpatient discharges dataset by Patient ID, Sex, Provider ID, Admission Date, and Discharge Date.
5. Remove the IPF inpatient discharges for patients who do not have Part A and Part B coverage at admission, during the entire stay, at discharge, and during the 30 days post-discharge.
6. Remove the IPF inpatient discharges that do not have a principal diagnosis of MDD, bipolar disorder, or schizophrenia using value sets containing ICD-9 codes for each of the disease conditions.
7. Remove the IPF inpatient discharges for patients who expired during the hospital stay or within 30 days of discharge.
8. Remove the IPF inpatient discharges for patients who do not have Part D coverage during the 30 days post-discharge.
9. Remove the IPF inpatient discharges for patients who were not discharged to home or home health.
10. Exclude IPF inpatient discharges with a secondary diagnosis of pregnancy or delirium.
11. Exclude IPF inpatient discharges having schizophrenia as the principal diagnosis with a secondary diagnosis of dementia.
12. Exclude IPF inpatient discharges with ECT or TMS during the hospital stay or within 30 days post-discharge.

Numerator:
1. Pull all Part D claims for the evidence-based medications used for the treatment of MDD, schizophrenia, and bipolar disorder.
2. Pull all Part A and Part B claims for antipsychotic long-acting injectables (LAIs) and add them to the Part D medication claims for schizophrenia and bipolar disorder.
3. Compare the medication claims to the denominator file of eligible IPF inpatient discharges and remove any claims that occur more than 2 days prior to the discharge date.
4. Determine which claims occur within the follow-up period (2 days prior to discharge through 30 days post-discharge) for each of the 3 disease conditions.
5. Total the denominator cases having at least one medication claim corresponding to the disease condition during the follow-up period.

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

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.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
This measure is not based on survey or patient-reported data.

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

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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Medicare administrative data from Parts A, B, and D claims.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

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

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health : Inpatient
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 because this is not a composite performance measure.

2. Validity – See attached Measure Testing Submission Form
Med_Continuation_nqf_testing_attachment-636174985400164068.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. (Do not remove prior testing information – include date of new information in red.)

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. (Do not remove prior testing information – include date of new information in red.)

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

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

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

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

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Medication Continuation Following Inpatient Psychiatric Discharge
Date of Submission: 12/16/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

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

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

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

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

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

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

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

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

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

Medicare administrative claims data

1.3. What are the dates of the data used in testing? January 1, 2013- January 31, 2015

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

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

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

The measure was developed and tested using Medicare files for all inpatient psychiatric facility (IPF) discharges that occurred between January 1, 2013 and December 31, 2014. The data include 380,861 discharges from 1,694 IPFs across the United States (Table 1.5-A). IPFs ranged in size from 4 to 771 inpatient beds. Approximately 70% of IPFs in this dataset were units within a larger hospital. The average number of discharges per freestanding IPF was approximstely 300 and the average per IPF unit was approximately 200.

Table 1.5-A. Distribution of Discharges by IPF Type (January 1, 2013 – December 31, 2014)

IPF Type	IPFs (N=1,694)	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
Freestanding	515	301.8	322.9	1	20	77	184	416	779	1,760
Unit	1,179	191.2	189.9	1	24	56	135	263	419	1,320
Overall	1,694	224.8	243.6	1	23	60	148	293	529	1,760

To inform the preliminary measure specifications, we conducted alpha testing, which consisted of medical record review in two IPFs at a large academic medical center in the southeast U.S.

To evaluate the validity of key elements in the claims data, we conducted similar medical record abstractions in seven additional IPFs. Test sites varied in size, type, and geographic location (Table 1.5-B).

Table 1.5-B. Characteristics of Test Sites

Study ID	State	Bed Size	Type	Teaching Facility	Type of Medical Record
1	WV	Large	Unit	Yes	EPIC
2	MI	Medium	Unit	Yes	McKesson
3	AZ	Medium	Freestanding	No	Paper Records
4	AZ	Large	Freestanding	No	Paper Records
5	MD	Large	Freestanding	Yes	Allscripts®
6	CA	Small	Unit	No	Cerner
7	LA	Large	Unit	Yes	Epic

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*)

This measure was developed for adult admissions to an IPF with a principal diagnosis of major depressive disorder (MDD), schizophrenia, or bipolar disorder. Eligible patients were enrolled in Medicare Parts A, B, and D during the admission and follow-up period. The final cohort includes 380,861 discharges. On average, 35% of discharges had a principal diagnosis of MDD, 40% of discharges had a principal diagnosis of schizophrenia, and 27% of discharges had a principal diagnosis of bipolar disorder (Table 1.6-A). When limiting to facilities with 75 or more cases during the measurement period (rationale provided in Section 2a.2), 30% of discharges had a principal diagnosis of MDD, 43% of discharges had a principal diagnosis of schizophrenia, and 27% of discharges had a principal diagnosis of bipolar disorder on average (Table 1.6-B). The patients in the claims data were 51% male, 84% under age 65, and 70% dually enrolled. The racial and ethnic groups represented were 72% white, 21% black, and 4% Hispanic.

Table 1.6-A. Distribution of Bipolar Disorder, MDD, and Schizophrenia Across IPFs

Condition	IPFs	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
MDD	1,651	34.8	19.0	0.8	11.7	21.4	32.5	45.7	61.4	100
Schizophrenia	1,655	40.2	19.9	0.6	15.2	25.7	38.0	52.7	67.4	100
Bipolar Disorder	1,658	27.3	11.8	1.0	14.3	20.0	26.1	33.3	40.6	100

Table 1.6-B Distribution of Bipolar Disorder, MDD, and Schizophrenia Across IPFs with Denominator ≥ 75

Condition	IPFs	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
MDD	1,182	29.5	14.7	0.8	11.2	19.3	28.7	38.6	48.1	91.3
Schizophrenia	1,184	43.1	17.3	0.6	23.1	30.7	40.8	54.0	67.2	96.1
Bipolar Disorder	1,184	27.4	9.5	1.0	15.7	21.1	26.9	33.3	39.7	76.3

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.

Most data analysis was conducted in claims data. As noted in Section 1.5, alpha testing data from medical record review at two sites helped to inform the measure specifications. Medical records for 166 discharges were abstracted by two clinicians.

The field testing that informed the validity of key data elements was conducted by two nurses at each facility. Each nurse abstracted medical records for 75 discharges each for a total of 150. Twenty percent of each nurse's discharges were randomly selected and assigned to the other nurse abstractor to assess the reliability of the nurse abstractions. Additionally, two clinicians per facility reviewed a sub-sample (10 percent) of the medical records of the 150 discharges to determine the validity of the principal diagnosis, based on information contained in the record. Fifty percent of each clinician's discharges were randomly selected and assigned to the other clinician abstractor to assess the reliability of the clinician abstractions. Reliability scores between the two clinicians were calculated.

At the start of testing, each test site received a one-hour training by HSAG on the abstraction instructions and process and a one-hour follow-up meeting after review of the first 10 medical records to provide clarifications, if needed.

The abstraction tool that was used by all field testing sites is provided in the measure technical report in the supplemental materials.

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

Not applicable because the measure is not risk adjusted or stratified.

2a2. RELIABILITY TESTING

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

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

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

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

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

To examine the reliability of the measure score, we utilized the approach proposed by Adams (2009) and Scholle et al. (2008) to assess measure precision in the context of the observed variability across IPFs. 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."

For this measure, the signal-to-noise ratio was calculated as a function of the variance between IPFs (signal) and the variance within an IPF (noise). Reliability was estimated using a beta-binomial model. This approach has two basic assumptions:

1. Each measured entity has a true pass rate, p, which varies; and,
2. The measured entity's score is a binomial random variable conditional on the measured entity's true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of 0.0 implies that all variation is attributed to measurement error (noise); whereas, a reliability of 1.0 implies that all variation is caused by a real difference in performance (across IPFs). 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; thus, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to discriminate performance between IPFs.

Using methodology described by Scholle et al. (2008), reliability estimates were computed separately, based on the mean denominator size for IPFs 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 IPFs in this population."

*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*)

A minimum denominator size of 75 discharges is needed to attain an overall reliability score of at least 0.7 (Table 2a.2.3-A), which is within acceptable norms and indicates sufficient signal strength to discriminate performance between facilities, using the method of mean denominator and volume categories. With a minimum denominator of 75 discharges, 1,184 IPFs (70%) have enough discharges within a two-year measurement period for public reporting. The removal of smaller facilities does not have an appreciable impact on the distribution of measure scores (Table 2a.2.3-B).

Table 2a.2.3-A. IPF Reliability and Assessment of Adequacy for Tests Conducted

	Minimum Denominator	# of IPFs N=1,694 (%)	Mean Rate (%) of IPFs	Reliability Score
Overall	75	1,184 (69.9)	78.0	0.77

Table 2a.2.3-B. Comparison of IPF Measure Score Distribution by Denominator Minimum

	# IPFs	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
Overall	1,694	78.0	11.1	0.0	66.7	73.6	79.6	84.4	88.3	100.0
Denominator ≥ 75	1,184	78.0	7.9	21.1	68.3	73.9	79.1	83.4	86.5	98.5

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 results indicate the measure score is reliable by adjusting the minimum case size for the denominator to require at least 75 cases during the measurement period. To increase the number of IPFs that have at least 75 cases during the measurement period, we recommend using a two-year measurement period.

2b2. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Critical data elements

Two psychiatrists reviewed 150 patients' medical records to ensure that the claims data are accurate in identifying several key data elements for calculating the measure. First, the clinicians recorded their assessment of the patient's principal discharge diagnosis based on information in the medical record. These findings were compared to the principal diagnoses in the claims. We evaluated the positive predictive value using the clinical assessment from the medical record as the "gold standard" because this shows how often a diagnosis in the claims agrees with the diagnosis from the medical record. A high positive predictive value indicates a high probability that a claim for a certain condition (e.g., schizophrenia) correctly predicts the principal discharge diagnosis in the medical record.

Next, at the seven test sites, abstractors were asked to indicate whether a prescription was provided at discharge. When an evidence-based prescription was not provided, they were asked to provide the rationale from the medical record to determine if additional exclusion criteria should be applied to the measure. The information on whether at least one prescription for an evidence-based medication was provided at discharge was compared to the numerator based on claims data. We evaluated the positive predictive value using the prescription at discharge as the "gold standard". The positive predictive value indicates that most patients who filled an evidence-based prescription during the follow-up period also received an evidence-based prescription from the IPF at discharge.

Finally, abstractors from the seven test sites were asked to record whether there was an indication in the medical record that medications had been dispensed to the patient free at discharge, as those medications would not appear in the claims data.

To ensure that the abstraction results were reliable, 10% of the cases were reviewed by both clinicians, and their results were compared to assess agreement.

Performance measure score

Measure scores were compared to three related measures:

1. Follow-Up After Hospitalization (7-Day)
2. Follow-Up After Hospitalization (30-Day)
3. IPF All-Cause Unplanned Readmission Measure

We tested the measure distributions for normality at each unit of analysis, selected the appropriate statistical test for the distribution, and assessed the significance of the correlation coefficient. We would expect the scores for the 7- and 30-day Follow-Up After Hospitalization measure to be positively correlated with the medication continuation scores because these are care coordination measures and higher scores indicate higher quality. We would expect the medication continuation scores to be negatively correlated with the all-cause unplanned readmission measure scores, because readmissions may indicate a lack of care coordination and higher scores on the readmission measure indicate lower quality.

Face validity of the measure score was assessed by the IPF Technical Expert Panel (TEP). Specifically, the TEP members were asked whether they agreed, disagreed, or were unable to rate the following statement:

> *The performance rating from the continuation of medication measure, as specified, represents an accurate reflection of facility-level rates of evidence-based medication continuation for MDD, schizophrenia, or bipolar disorder following discharge from an IPF.*

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

Critical data elements

The positive predictive value of the claims data was 97% (921/945) (Table 2b2.3-A). The positive predictive values were similar across all three conditions, with 98% (289/294) for MDD, 98% (328/335) for schizophrenia, and 96% (304/316) for bipolar disorder.

Table 2b2.3-A. Agreement Between Medical Record and Claims for Diagnoses

	Diagnosis In Medical Record	Diagnosis Not in Medical Record	Total
MDD			
MDD in claims	289	5	294
No MDD in claims	6	0	6
Total MDD	295	5	300
Schizophrenia			
Schizophrenia in claims	328	7	335
No schizophrenia in claims	9	0	9
Total schizophrenia	329	7	344
Bipolar Disorder			
Bipolar disorder in claims	304	12	316
No bipolar disorder in claims	3	0	3
Total bipolar disorder	307	12	319
Total Overall	939	24	963

During the medical record review at the 7 test sites, 92% (873/945) of cases were prescribed an evidence-based medication at discharge (Table 2b2.3-B). Among the patients who were not prescribed an evidence-based medication, the majority of reasons identified by the medical record abstractors indicated quality deficits. For example, 61% of the cases without an evidence-based medication at discharge had medications prescribed that were not indicated for the principal discharge diagnosis, 11% did not have any medications prescribed, and 5% were clearly the result of medical errors. No reason was identified by the abstractors for 9% of the cases, which could also indicate potential quality deficits. The remaining cases do not represent quality deficits but do indicate opportunities for improvement in cases where prescriptions could have been provided in addition to medications dispensed at discharge or could have been provided to patients who declined pharmacotherapy because the patient may decide differently and want to continue pharmacotherapy after leaving the IPF.

When comparing numerator positive cases from the claims data to the medical record, the positive predictive value was 96% (622/646) as calculated from Table 2b2.3-B.

Table 2b2.3-B. Comparison of Medications Prescribed at Discharge to Fills During the Follow-Up Period in Claims Data

	Evidence-Based Prescription at Discharge	No Evidence-Based Prescription at Discharge	Total
Numerator Positive	622	24	646
Numerator Negative	251	48	299
Total	873	72	945

The medical record review found that there were few discharges where the facility provided medications to patients at discharge. Among those discharges, some of the medications provided were filled for the patient through an outpatient pharmacy and appeared in the claims data.

Performance measure score

Results of the analysis for correlations of medication continuation scores with the three conceptually related Inpatient Psychiatric Facility Quality Reporting (IPFQR) measures are included in Table 2b2.3-C. The medication continuation scores were moderately correlated with the scores for 7- and 30-day follow-up after hospitalization for mental illness scores as expected (ρ = 0.34 and 0.43). The medication continuation scores were negatively correlated with readmission scores as expected (ρ = -0.26). All correlations are statistically significant at p-value < 0.0001.

After reviewing these results and the proposed measure specifications, all of the 10 TEP members who were present for the face validity vote agreed that the measure score had face validity.

Table 2b2.3-C. Performance Measure Score Correlation

Measure	IPFs	Correlation
Follow-Up After Hospitalization 7-day (7/1/2014 – 6/30/2015)	1,145	0.34312
Follow-Up After Hospitalization 30-day (7/1/2014 – 6/30/2015)	1,145	0.43065
IPF All-Cause Unplanned Readmission Measure (Observed) (1/1/2013 – 12/31/2014)	1,184	-0.26059

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

Critical data elements

The medical record review in the two initial test sites confirmed that the principal discharge diagnoses in the administrative claims data are a valid source for identifying the primary cause of admission to the IPF.

The medical record review from the additional 7 test sites confirmed that the construct of medication continuation is valid for assessing IPF quality because most patients who filled a prescription during the follow-up period received a prescription from the IPF at discharge. A quality deficit was identified for most patients who were not provided a prescription for an evidence-based medication at discharge so no additional exclusion criteria were applied to the measure as the result of this analysis.

Finally, the medical record review at the seven test sites confirmed that the claims data are valid for identifying all prescription fills in this patient population because medications provided at discharge were filled using the patient's insurance, which would appear in the claims data. We anticipate that free medications are provided to the patient population for this measure less frequently because all patients included in the measure denominator are enrolled in Medicare Part D. Low-income Medicare patients can receive assistance with co-pays, and patients who are dually enrolled in Medicaid (70% of this cohort) receive additional assistance covering the costs of medications that are not covered by Medicare. Notes from the medical record abstractors indicate that all of the medications provided at discharge were for 30-day supplies or less. Therefore, the patients who received medications at discharge on Day 0 would need to fill a prescription for an evidence-based medication before the end of the 30-day follow-up period to avoid gaps in treatment. Those fills would also appear in the claims data.

Performance measure score

The moderate strength of the correlations, conceptually supported directionality, and unanimous face validity assessment add further support that the measure is valid as specified.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

All exclusion analyses were conducted using Medicare claims data from inpatient psychiatric stays at IPFs where the patients were discharged alive with Parts A, B, and D enrollment during the follow-up period.

1. **Electroconvulsive therapy (ECT)**
   We compared the medication continuation rates of patients with ECT during the admission or follow-up period to those of patients without ECT during the admission or follow-up period. We also conducted a medical record review to evaluate whether evidence-based medications were prescribed at discharge to patients who received ECT or a recommendation for ECT.

2. **Transcranial magnetic stimulation (TMS)**
   We compared the medication continuation rates for patients with TMS during the admission or follow-up period to those of patients without TMS during the admission or follow-up period.

3. **Pregnancy**
   We compared the medication continuation rates for patients who were pregnant during the admission to those of patients who were not pregnant during the admission.

4. **Secondary diagnosis of delirium**
   We compared the medication continuation rates for patients with delirium during the admission to those of patients without delirium during the admission.

5. **Principal diagnosis of schizophrenia with secondary diagnosis of dementia**
   Antipsychotics may be contraindicated for patients with dementia. Antipsychotics are included in the numerator for schizophrenia and bipolar disorder. However, alternative pharmacotherapies are available for bipolar disorder that meet the numerator criteria, so we only compared the medication continuation rates for patients with a principal diagnosis of schizophrenia and a secondary diagnosis of dementia to those of patients with no dementia.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

1. **ECT**

Table 2b3.2-A. Frequency of ECT During or After the Index Admission

Principal Condition	All IPF Admissions		ECT During Admission Or Follow-Up Period			No ECT During Admission Or Follow-Up Period		
	Frequency	% Rx	Frequency	% Total	% Rx	Frequency	% Total	% Rx
MDD	139,355	71.7	7,414	5.3	76.3	131,941	94.7	71.4
Schizophrenia	217,417	75.6	3,086	1.4	77.3	214,331	98.6	75.5
Bipolar disorder	132,376	75.5	4,474	3.4	74.6	127,902	96.6	75.6
Overall	489,148	74.5	14,974	3.1	76.0	474,174	96.9	74.4

2. **TMS**

Table 2b3.2-B. Frequency of TMS During the Stay or After the Index Admission for MDD, Schizophrenia, or Bipolar Disorders

Principal Condition	All IPF Admissions		TMS During Admission Or Follow-Up Period			No TMS During Admission Or Follow-Up Period		
	Frequency	% Rx	Frequency	% Total	% Rx	Frequency	% Total	% Rx
Overall	489,148	74.5	76	0.0	76.3	489,072	100.0	74.5

3. **Pregnancy**

Table 2b3.2-C. Follow-Up Rates for Patients Who Are and Are Not Pregnant

Condition	All IPF Admissions	Pregnant	Not Pregnant

	Frequency	% Rx	Frequency	% Total	% Rx	Frequency	% Total	% Rx
MDD	139,355	71.7	59	0.0	59.3	139,296	99.9	71.7
Schizophrenia	217,417	75.6	138	0.1	59.4	217,279	99.9	75.6
Bipolar disorder	132,376	75.5	134	0.1	61.9	132,242	99.9	75.5
Overall	489,148	74.5	331	0.1	60.4	488,817	99.9	74.5

4. Secondary diagnosis of delirium

Table 2b3.2-D. IPF Admissions with Secondary Delirium Diagnosis

	All IPF Admissions		Delirium			No Delirium		
Principal Condition	Frequency	% Rx	Frequency	% Total	% Rx	Frequency	% Total	% Rx
MDD	139,355	71.7	3,420	2.5	66.5	135,935	97.5	71.8
Schizophrenia	217,417	75.6	3,837	1.8	71.9	213,580	98.2	75.6
Bipolar disorder	132,376	75.5	2,385	1.8	73.2	129,991	98.2	75.6
Overall	489,148	74.5	9,642	2.0	70.3	479,506	98.0	74.5

5. Principal diagnosis of schizophrenia and secondary diagnosis of dementia

Table 2b3.2-E. IPF Admissions with Principal Diagnosis of Schizophrenia and Secondary Diagnosis of Dementia

	All IPF Admissions		Secondary Dementia			No Dementia		
Principal Condition	Frequency	% Rx	Frequency	% Total	% Rx	Frequency	% Total	% Rx
Schizophrenia	217,417	75.6	6,971	3.2	65.3	210,446	96.8	75.9

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

1. ECT

ECT procedures are used as a form of treatment in the IPF patient population (3.1%), and many patients receiving ECT filled evidence-based medications during the follow-up period. However, given that ECT may be used as an alternative when patients fail pharmacotherapy and that the medical record review showed that patients receiving ECT did not always receive an evidence-based prescription, the TEP and workgroup recommended the exclusion from the denominator of patients receiving ECT during the index admission or follow-up period.

2. TMS

TMS is a newer procedure and is still rare. Many patients receiving TMS also filled evidence-based medications during the follow-up period. However, since TMS may be used as an alternative when patients fail pharmacotherapy, the TEP and workgroup recommended the exclusion of patients receiving TMS during the index admission or follow-up period from the denominator.

3. **Pregnancy**
   Pregnancy was rare in this patient population (0.1%). The results showed that pregnant patients had empirically lower rates of filling evidence-based medications within 30 days of discharge than patients who were not pregnant (60.4% compared to 74.5%), which supports the TEP and workgroup recommendations to exclude from the denominator. Therefore, we excluded pregnant patients from the measure.

4. **Secondary diagnosis of delirium**
   Patients with secondary diagnoses of delirium are rare (2.0%). The results showed that patients with delirium had empirically lower rates of filling evidence-based medications within 30 days of discharge than patients without delirium (70.3% compared to 74.5%), which supports the TEP and workgroup recommendations to exclude from the denominator. Therefore, we excluded patients with delirium from the measure.

5. **Principal diagnosis of schizophrenia with secondary diagnosis of dementia**
   Patients with schizophrenia and secondary diagnoses of dementia were rare (3.2%). The results showed that patients with schizophrenia and a secondary diagnosis of dementia had empirically lower rates of filling evidence-based medications within 30 days of discharge than patients without dementia (65.3% compared to 75.9%), which supports the TEP and workgroup recommendations to exclude from the denominator. Therefore, we excluded patients with schizophrenia and a secondary diagnosis of dementia from the measure.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

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

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

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

2b4.1.1 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 because the measure is not risk adjusted.

2b4.2. 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 because this is a process measure.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not applicable because this measure is not risk adjusted.

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

Not applicable because this measure is not risk adjusted.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable because this measure is not risk adjusted.

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

Not applicable because this measure is not risk adjusted or stratified.

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

If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Not applicable because this measure is not risk adjusted.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Not applicable because this measure is not risk adjusted.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Not applicable because this measure is not risk adjusted.

2b4.9. Results of Risk Stratification Analysis:

Not applicable because this measure is not stratified.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)

Not applicable because this measure is not risk adjusted or stratified.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*) Not applicable

Not applicable because this measure is not risk adjusted or stratified.

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

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

To evaluate whether there is currently a performance gap and variation in performance across facilities, we applied all inclusion and exclusion criteria to calculate facility-level measure scores. We observed the distribution of medication continuation rates and the difference between IPFs in the 90th percentile of performance and IPFs in the 10th percentile. To identify statistically significant differences in performance, we calculated 95% confidence intervals (95% CI) around the measure scores for each IPF and compared the 95% CI to the national medication continuation rate across all IPFs. If the confidence intervals did not overlap with the national medication continuation rate, the difference was considered statistically significant.

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

An analysis of 2013-2014 Medicare claims data indicated performance varied between high- and low-performing facilities across more than 1,600 IPFs for each of the three diagnoses (Table 2b5.2-A). For the combined measure score, there is about a 22 percentage point difference between the 10th and 90th percentiles (66.7%–88.3%) and a median score of 79.6%.

Table 2b5.2-A. Distribution of Facility Performance

Diagnosis	# IPFs	Mean	SD	Min	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	Max
MDD	1,651	75.5	13.9	0.0	60.0	69.6	77.1	83.3	89.7	100.0
Schizophrenia	1,655	79.1	15.3	0.0	63.6	73.1	81.5	87.9	95.5	100.0
Bipolar disorder	1,658	78.3	14.4	0.0	63.9	72.5	80.0	86.4	93.5	100.0
Overall	1,694	78.0	11.1	0.0	66.7	73.6	79.6	84.4	88.3	100.0

About 24% of facilities had medication continuation rates that were statistically better than the national rate, and about 13% of facilities had medication continuation rates that were statistically worse than the national rate (Table 2b5.2-B).

Table 2b5.2-B. Distribution of IPFs Compared to the National Medication Continuation Rate

Performance Categorization	Count IPFs	Percent IPFs
Total IPFs	1,694	100.0
Better than national rate	399	23.6
No different than national rate	572	33.8
Worse than national rate	213	12.6
Fewer than 75 discharges during the performance period	510	30.1

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
The results indicate ample room for improvement and meaningful differences in the quality of care between the highest and lowest performing facilities.

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

Not applicable because there is only one set of specifications.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

Not applicable because there is only one set of specifications.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

Not applicable because there is only one set of specifications.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

Not applicable because this measure is based on claims data.

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

Not applicable because this measure is based on claims data.

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

Not applicable because this measure is based on claims 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.**
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**.
No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
Not applicable because this measure is based on administrative claims data.

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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
Data used in the calculation of this measure are obtained from administrative claims, which are routinely, reliably, and securely collected for billing purposes. We do not anticipate any feasibility or implementation issues related to data collection for this 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)*.
There are no fees, licensing, or other requirements associated with the use of this 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 Not in use	

4a.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

This measure is not currently in use.

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

New measure under development.

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

This measure has been submitted to the Measures Under Consideration (MUC) list to be reviewed by the Measure Applications Partnership (MAP) for use in the IPFQR Program.

Improvement

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

4b. 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 new and has not been implemented yet so trends in performance are not available. By calculating the facility-level medication continuation scores in Medicare FFS claims data and providing them to facilities, CMS aims to encourage quality improvement, specifically relating to stronger care transitions to outpatient settings. Literature about continuation of medication has identified effective interventions that facilities can employ to improve medication adherence among patients discharged from an IPF (Douaihy, Kelly, & Sullivan, 2013; Haddad, Brain, & Scott, 2014; Hung, 2014; Kasckow & Zisook, 2008; Lanouette, Folsom, Sciolla, & Jeste, 2009; Mitchell, 2007; Sylvia et al., 2013). Examples of these interventions include patient education, shared decision-making, and text-message reminders. We envision the addition of this measure to the suite of measures for IPFs would help to create a comprehensive picture of the quality of care patients receive at those facilities.

*Douaihy, A. B., Kelly, T. M., Sullivan, C. (2013). Medications for substance use disorders. Social Work in Public Health, 28(3-4), 264-278. doi: 10.1080/19371918.2013.759031
*Haddad, P. M., Brain, C., & Scott, J. (2014). Nonadherence with antipsychotic medication in schizophrenia: Challenges and management strategies. Patient Related Outcome Measures, 5, 43-62. doi: 10.2147/PROM.S42735
*Hung, C. I. (2014). Factors predicting adherence to antidepressant treatment. Current Opinion in Psychiatry, 27(5), 344-349. doi: 10.1097/yco.0000000000000086
*Kasckow, J. W., & Zisook, S. (2008). Co-occurring depressive symptoms in the older patient with schizophrenia. Drugs & Aging, 25(8),631-647.
*Lanouette, N. M., Folsom, D. P., Sciolla, A., Jeste, D. V. (2009). Psychotropic medication nonadherence among United States Latinos: A comprehensive literature review. Psychiatric Services (Washington, DC), 60(2), 157-174. doi: 10.1176/appi.ps.60.2.157
*Mitchell, A. J. (2007). Understanding medication discontinuation in depression. BMedSci Psychiatric Times, 24(4).
*Sylvia, L. G., Hay, A., Ostacher, M. J., et al. (2013). Association between therapeutic alliance, care satisfaction, and pharmacological adherence in bipolar disorder. Journal of Clinical Psychopharmacology, 33(3), 343-350. doi: 10.1097/JCP.0b013e3182900c6f

4c. Unintended Consequences

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

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

The measure is not yet implemented.

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

The measure is not yet implemented.

4d1.1. **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.

Individual performance results and assistance with interpretation will be provided to IPFs if the measure is implemented.

4d1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**

Individual performance results and assistance with interpretation will be provided to IPFs if the measure is implemented.

4d2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**

Describe how feedback was obtained.

Feedback from IPFs will be provided during endorsement maintenance if the measure is implemented.

4d2.2. **Summarize the feedback obtained from those being measured.**

Feedback from IPFs will be provided during endorsement maintenance if the measure is implemented.

4d2.3. Summarize the feedback obtained from other users
Feedback from other users will be provided during endorsement maintenance if the measure is implemented.

4d.3. Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.
Information on potential measure revisions based on IPF and other user feedback will be provided during endorsement maintenance.

5. Comparison to Related or Competing Measures

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

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

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

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

5a. Harmonization 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.
Not applicable because there are no related measures.

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 because there are no competing measures.

Appendix

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

Attachment **Attachment:** 161216_Methodology_Report_Med_Continuation_for_NQF.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services, Contracting Officer's Representative (COR)
Co.2 **Point of Contact:** Vinitha, Meyyur, vinitha.meyyur@cms.hhs.gov, 410-786-8819-
Co.3 **Measure Developer if different from Measure Steward:** Health Services Advisory Group, Inc. (HSAG)
Co.4 **Point of Contact:** Megan, Keenan, mkeenan@hsag.com, 616-425-1997-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Inpatient Psychiatric Facility (IPF) Outcome and Process Measure Development and Maintenance Technical Expert Panel (TEP):

Alisa Busch, MD, MS
Director, Integration of Clinical Measurement & Health Services Research
Chief, Health Services Research Division, Partners Psychiatry and Mental Health
Assistant Professor of Psychiatry and Health Policy, Harvard Medical School

Kathleen Delaney, PhD, PMH-NP, RN
Professor, Rush College of Nursing

Jonathan Delman, PhD, JD, MPH
Assistant Research Professor, Systems and Psychosocial Advances Research Center, University of Massachusetts Medical School

Frank Ghinassi, PhD, ABPP
Vice President, Quality and Performance Measurement, Western Psychiatric Institute and Clinic
Associate Professor in Psychiatry, University of Pittsburg

Eric Goplerud, PhD
Senior Vice President, Director of Public Health Department, NORC at the University of Chicago

Geetha Jayaram, MD
Associate Professor, Schools of Medicine, Health Policy and Management and the Armstrong Institute for Patient Safety, Johns Hopkins University

Charlotte Kauffman, MA, LCPC
Service Systems Coordinator, State of Illinois-Division of Mental Health

Tracy Lenzini, BS
Executive Director, Grand Traverse Health Advocates

Kathleen McCann, RN, PhD

Director of Quality and Regulatory Affairs, National Association of Psychiatric Health Systems

Gayle Olano-Hurt, MPH, CPHQ, PMC
Director Data Management, Outcomes Measurement & Research Administration, Sheppard Pratt Health System

Mark Olfson, MD, MPH
Professor of Psychiatry, Columbia University Medical Center Department of Psychiatry; New York State Psychiatric Institute

Irene Ortiz, MD, MSW
Medical Director, Molina Healthcare of New Mexico

Thomas Penders, MS, MD, DLFAPA
Medical Director, Inpatient Psychiatry, Vident Medical Center
Associate Professor, Brody School of Medicine Department of Psychiatry, East Carolina University

Lucille Schacht, PhD
Senior Director, Performance and Quality Improvement, National Association of State Mental Health Program Directors Research Institute, Inc.

Lisa Shea, MD
Medical Director, Butler Hospital
Thomedi Ventura, MS, MSPH
Program Evaluator, Telligen

Elvira Ryan, MBA, BSN, RN
Associate Project Director, Division of Healthcare Quality Evaluation, The Joint Commission

Measure Workgroup Members:
TEP Members:
Frank Ghinassi, PhD
Geetha Jayaram, MD
Charlotte Kauffman, MA
Kathleen McCann, PhD, RN
Gayle Olano-Hurt, MPH
Thomedi Ventura, MSPH

UF Members:
Regina Bussing, MD
Professor and Chair, Department of Psychiatry, University of Florida College of Medicine

Mathew Nguyen, MD
Assistant Professor and Medical Director, Department of Psychiatry, University of Florida College of Medicine

Gary Reisfield, MD
Associate Professor, Department of Psychiatry, University of Florida College of Medicine

Almut Winterstein, PhD, RPh, FISPE
Professor and Chair, Pharmaceutical Outcomes and Policy, University of Florida College of Medicine

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** Not applicable. Ad.7 **Disclaimers:** None.
Ad.8 **Additional Information/Comments:** None.

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 #: 3207
Measure Title: Medication Reconciliation on Admission
Measure Steward: Centers for Medicare & Medicaid Services, Contracting Officer's Representative (COR)
Brief Description of Measure: The average completeness of the medication reconciliation process within 48 hours of admission to an inpatient facility.
Developer Rationale: The measure was constructed as a composite because of its scoring methodology that first computes scores for each of its three components on the facility level and then averages these scores into a single score. While each component is necessary to describe a single construct, medication reconciliation, presentation of the measure as composite allows close examination of the various parts that define the final single score. Thus, the measure architecture is based on the following considerations:

Alignment with Existing Best Practices: The three components in the measure align with the three core components of Medication Reconciliation on Admission specified by The Joint Commission NPSG.03.06.01.

Simplification of Measure Scoring: A subset of the data elements abstracted for this measure as defined by the data dictionary are used to calculate facility scores. The items in the subset are referred to as scoring elements. Scoring elements in Component 1 are assessed at the medical record level and scoring elements in Components 2 and 3 are assessed at the medication level. The average number of PTA medications per patient varies dramatically across patients as well as across facilities (based on facility case mix, especially related to age). Aggregation of medication information into a single facility-level score for Components 2 and 3 ensures that the scoring elements in Components 2 and 3 contribute consistently to the final measure scores across facilities regardless of the number of medications that were abstracted per record.

Numerator Statement: This measure does not have a traditional numerator. The numerator is a facility-level score of the completeness of the medication reconciliation process within 48 hours of admission. This score is calculated by averaging the scores of the three components of the medication reconciliation process. The components include:

1) Comprehensive prior to admission (PTA) medication information gathering and documentation
2) Completeness of critical PTA medication information
3) Reconciliation action for each PTA medication

Denominator Statement: The denominator for the composite measure includes admissions to an inpatient facility from home or a non-acute setting with a length of stay greater than or equal to 48 hours.
Denominator Exclusions: This measure does not have any denominator exclusions.

Measure Type: Composite
Data Source: Other, Paper Records
Level of Analysis: Facility

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.

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.
- A 2012 systematic review of hospital-based medication reconciliation practices concluded that it led to reduction in medication discrepancies, potential adverse drug events, and adverse drug events. There was inconsistent reduction in post-discharge healthcare utilization.
    - Key aspects included targeting the intervention to high-risk population.
    - The review did not discriminate whether the reconciliation happened at admission, transfer between units, or discharge.
    - Evidence not graded overall. Out of 26 studies, 6 were rated as good quality, 5 as fair, and 15 as poor.
    - No specific recommendations made.
    - The developer identified 10 relevant newer studies. Of these, 7 found reductions in medication discrepancies or lower rates of adverse drug events.
        - One study specific to the inpatient psychiatric facility setting found that updating and standardizing medication reconciliation processes increased the accuracy of medications from 45% to 80%.
- In addition, while not considered evidence *per se*, the developer provides The Joint Commission National Patient Safety Goals for hospitals (effective 2017) includes a goal to "maintain and communicate accurate patient medication information." Three aspects of this goal, which map to the 3 components of this measure, include:
    - "Obtain information on the medications the patient is currently taking when he or she is admitted to the hospital or is seen in an outpatient setting;"
    - "Define the types of medication information to be collected in non–24-hour settings and different patient circumstances;" and
    - "Compare the medication information the patient brought to the hospital with the medications ordered for the patient by the hospital in order to identify and resolve discrepancies."
- The developer also mentions a consensus statement from the Society of Hospital Medicine, but does not provide further information about this statement.

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence relevant enough considering measure focus on reconciliation at admission only?*

Guidance from the Evidence Algorithm

Composite measure based on systematic review (Box 3)→QQC presented (Box 4)→Quantity: moderate; Quality: moderate (11 of 26 good or fair); Consistency: high (Box 5b)→Moderate

The highest possible rating is 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.

- Nine inpatient psychiatric facilities (IPFs) from eight different states were used to perform the field testing of the measure. Both freestanding facilities and hospital-based units of various sizes and with different types of medical records were included. The sample included a total of 900 admissions from 1/4/2013 through 8/17/2016.
- The developer provides scores by facility for each of the 3 measure components (and their associated data elements).
    - Component 1: Comprehensive prior to admission (PTA) medication information gathering and documentation
    - Component 2: Completeness of critical PTA medication information
    - Component 3: Reconciliation action for each PTA medication
- The final overall score is calculated for each facility as the average of the scores for each of the 3 components.

Final Scores for each component and overall

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Component 1	79.0	32.4	86.8	81.4	37.8	61.5	60.7	79.2	55.3	63.8	32.4, 86.8
Component 2	80.7	89.7	96.1	96.0	76.9	74.0	85.7	76.7	74.2	83.3	74.0, 96.1
Component 3	74.2	76.3	57.5	98.8	63.5	14.0	78.0	99.5	23.3	65.0	14.0, 99.5
Overall Score	78.0	66.1	80.1	92.1	59.4	49.8	74.8	85.1	50.9	70.7	49.8, 92.1
95% CI	76.3, 79.6	64.0, 68.2	77.8 82.5	90.7, 93.4	56.7, 62.1	48.0, 51.7	72.3, 77.2	83.7, 86.6	48.7, 53.1	N/A	N/A

Disparities

- The developer provides data on performance during field testing stratified by demographic characteristic for the 900 admissions across the 9 facilities. This included 4,277 drugs as part of the reconciliation process.
- The developer notes that inferences are limited due to the small sample size. They also state that multivariate analysis would be needed to understand the relationships of the various demographic characteristics given the very diverse demographic distributions across the nine facilities, but that this is impractical given the small number of facilities examined.

Characteristic	Score
Gender	
Male	68.3
Female	69.1
Age	
0-18	68.3

19-24	68.4
25-34	71.3
35-44	75.7
45-54	74.7
55-64	76.9
>65	60.5
Missing	84.2
Race	
White	69.9
Black	65.5
Other	62.5
Ethnicity	
Hispanic	75.7
Non-Hispanic	70.8
Unknown	52.2
Insurance	
Medicaid	72.4
Medicare	66.0
Medicare + Medicaid	71.4
Other	69.4

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**

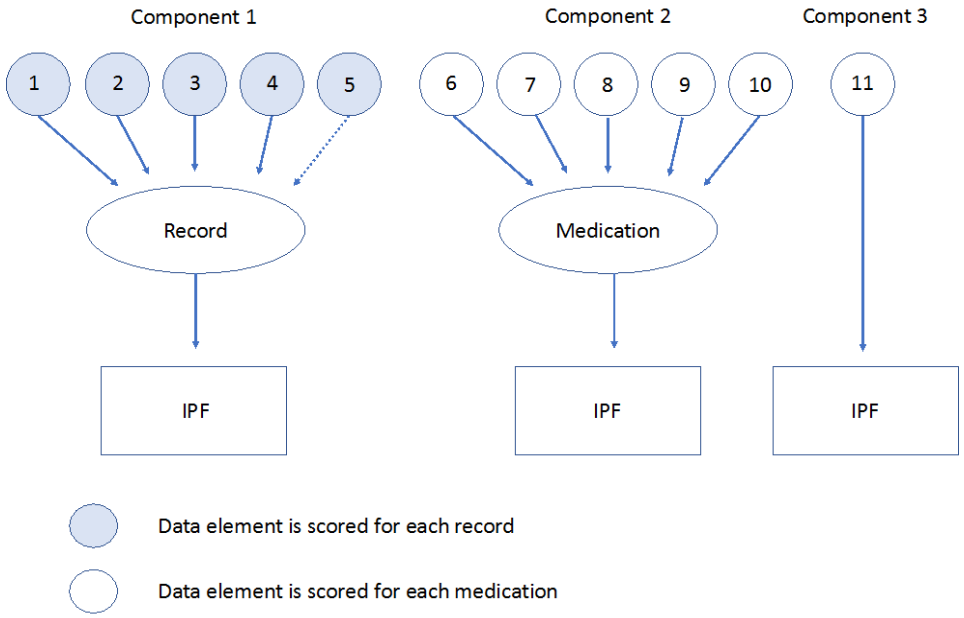
1c. Composite – Quality Construct and Rationale

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- The quality construct is described.
  - Overall, the measure calculates the average completeness of the medication reconciliation process within 48 hours of admission to an inpatient facility.
  - The three components (comprising 11 measure elements) align with the goals of The Joint Commission's National Patient Safety Goals.
  - Developer identifies the quality gap by citing cites studies related to discrepancies in patients' prescription medication histories, and those discrepancies' relationship to medication errors and adverse drug events.
- The 3 components are:
  - Component 1: Comprehensive prior to admission (PTA) medication information gathering and documentation
  - Component 2: Completeness of critical PTA medication information
  - Component 3: Reconciliation action for each PTA medication
- The rationale for the composite notes that component 1 is assessed at the medical record level while components 2 and 3 are assessed at the medication level. The developer states "Aggregation of medication information into a single facility-level score for Components 2 and 3 ensures that the scoring elements in

Components 2 and 3 contribute consistently to the final measure scores across facilities regardless of the number of medications that were abstracted per record."

- The aggregation and weighting of the measure are described for each component.
  - A facility-level score is calculated for each component and is based on individual scoring elements.
  - The overall facility-level score is the average of each of the three facility-level components. Each of the 3 components are weighted equally.
  - The developer provides a sensitivity analysis for determination of the final score composition.
- Measure Scoring Methodology:
  - The developer states that their methodology accommodates differing units of analysis for the 11 scoring elements and emphasizes the importance of the actual reconciliation action (Component 3). They state "We chose this methodology together with the measure development work group and the TEP after review of various alternatives and comparison of averages and ranges of the resulting scores across facilities. "
  - The developer presents explanation of measure scoring methodology, and gives a summary in the form of the figure below.



- The developer provides the individual scores by IPF and by data element for Component 1, Component 2, and Component 3, as well as a final overall score.

- The developer notes that in field testing, component 3 called for resolution of PTA medication discrepancies within 24 hours. They modified the specifications (to resolution within 48 hours) after the field testing. The technical expert panel agreed with this change. They added that this change was supported by a sensitivity analysis.

Questions for the Committee:

- *Are the quality construct and a rationale for the composite explicitly stated and logical?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*

Preliminary rating for composite quality construct and rationale: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

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

1a. Evidence to Support Measure Focus

Comments:
**There is evidence to support the measure focus is weak based on the literature.
**This composite measure was developed to look at rates of medication reconciliation completed at admission. The evidence cited suggests medication reconciliation results in fewer adverse medication events. The evidence is relevant, though reconciliation should be looked at upon discharge as well.
**Yes. The measure directly is related to the task of medication reconciliation. Evidence in 7 out of 10 studies demonstrated lower rates of discrepancies and adverse effects.

1b. Performance Gap
Comments:
**Data was provided for a performance gap and population subgroups.
**There is a performance gap. Reconciliation rates are lower with younger patients.
**Medication reconciliation is important especially with behavioral health admission where the reliability of information given can vary based on many factors. Disparities would be based on the source of info for meds prior to admission. Without verification from a reliable source, patient self report will be unreliable. Data source needed to be address in the analysis which is components 2 - 3.

1c. Composite Performance Measure
Comments:
**There is moderate quality constructs for the performance measures.
**The constructs are stated and are logical. The aggregation and weighing are explained and are logical.
**This appears to be a complex relationship and rational for the weighing of the rules for scoring. Not clear on the logic.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications **Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures**
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. **Data source(s):** Abstractions from paper records and electronic health records. **Specifications:** • Each component is scored at the facility level. • A higher score indicates better quality. • The numerator for this composite measure reflects the completeness of the medication reconciliation process, and is an average of the 3 components. Details are provided for criteria for the elements of each component. • The denominator for this composite measure includes admissions to an inpatient facility from home or a non-acute setting with a length of stay greater than or equal to 48 hours. o Transfer from another inpatient facility or inpatient unit are not included as medication reconciliation would have been completed by the facility or unit that admitted the patient. o Long-term care facilities and emergency departments are included as they are considered non-acute care settings. • There are no exclusions. • The measure is not risk adjusted. • A detailed calculation algorithm is provided. • The developer provides supplemental materials that include a data dictionary and measure information form that provide instructions for abstracting the data for the measure. • The developer notes that a structured chart abstraction tool with operational data definitions was developed in Excel for field testing. Prior to implementation, the measure developer states they will provide a finalized abstraction tool.

Questions for the Committee:

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*
- *Are there any questions about how the abstraction tool might be modified?*

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☐ **Data element** ☒ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data element reliability
  - Two abstractors at each IPF completed data ascertainment for all measure elements using a random subset of about 20 records (total subsample of 175 records).
  - They used a structured medical record abstraction tool (developed in Microsoft Excel) to collect data and calculate the measure score.
  - Inter-rater reliability (IRR) between the two abstractors at each site and for each measure data element was assessed using percent overall agreement and Cohen's Kappa statistic.
- Measure score reliability – a signal to noise analysis was performed.
  - Testing was based on a random sample of 100 charts from each of the 9 inpatient psychiatric facilities (IPF) (900 total admissions). The developer states "The final sampling strategy will be aligned with the current requirements of the IPFQR program with considerations to minimize the abstraction burden for facilities."
  - The sample population included both pediatric and adult admissions, with no restriction on insurance type.
  - The developer notes that in field testing, component 3 called for resolution of PTA medication discrepancies within 24 hours. They modified the specifications (to resolution within 48 hours) after the field testing. The technical expert panel agreed with this change. They added that this change was supported by a sensitivity analysis.

Results of reliability testing

- Data element reliability results

Percent of Agreement and Cohen's Kappa for measure score elements

Data Elements	All Records/ Medications	Agreed	% Agreement	Cohen's Kappa (pooled)^
Component 1				0.66 (0.49, 0.83)
Designated Medication Reconciliation Form/Area	175	166	94.9%	0.67 (0.02, 1.00)
Patient Source	175	134	76.6%	**0.20 (-0.46, 0.85)**
Health System Source	175	138	78.9%	0.52 (0.11, 0.94)

PTA Medication List Contains All H&P Medications	175	146	83.4%	0.57 (0.14, 1.00)
At least one medication is on PTA Medication List*	175	168	96.0%	0.88 (0.61, 1.00)
PTA Medication List Reviewed by Prescriber within 24 hours of Admission (for records with 0 medications)	42	39	92.9%	**0.22 (-1.00, 1.00)**
Component 2				0.71 (0.61, 0.81)
Number of Medications on PTA Medication List*	790	701	88.7%	/#
Medication Name	701	701	100.0%	/#
Medication Route	701	695	99.1%	0.91 (0.66, 1.00)
Medication Dose	701	693	98.9%	0.89 (0.61, 1.00)
Medication Frequency	701	685	97.7%	0.67 (0.25, 1.00)
Last Time Medication Taken	701	633	90.3%	0.59 (0.33, 0.84)
Component 3				0.62 (0.36, 0.88)
Medication Reconciliation Action within 24 hours of Admission	701	631	90.0%	0.62 (0.36, 0.88)
Total Score			91.3%	0.73 (0.66, 0.80)

*Added for purposes of reliability testing; not included in the measure score

cannot be calculated because of data structure (e.g., no disagreement or no variation in one category)

^ For simplicity and computational efficiency, we used the normal distribution formula to establish confidence intervals. The confidence intervals based on standard normal distribution may generate upper limits smaller than -1.00 or greater than 1.00, which were adapted to -1.00 and 1.00, respectively.

Cohen's Kappa within facilities

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Pooled Kappa
Type	Unit	Unit	FS	FS	FS	Unit	Unit	FS	FS	
Bed Size	70	28	90	75	322	12	38	24	168	
Data Source	EPIC	McKesson	Paper Medical Records	Paper Medical Records	Allscripts®	Cerner	EPIC	Netsmart TIER® CareRecord ™	Cerner	
Total Score (Kappa)	0.87 (0.84, 0.91)	0.71 (0.63, 0.78)	**0.42 (0.28, 0.55)**	0.70 (0.58, 0.83)	**0.42 (0.30, 0.53)**	0.83 (0.79, 0.87)	**0.55 (0.44, 0.66)**	0.92 (0.87, 0.96)	0.99 (0.98, 1.00)	0.73 (0.66, 0.80)

FS = free standing facility

- NOTE: Kappa values range between 0 and 1 and are interpreted as degree of agreement beyond chance. A common scale is used to interpret Kappa statistics: 0.01–0.20 is considered slight agreement; 0.21– 0.40 is fair agreement; 0.41–0.60 is moderate agreement; 0.61–0.80 is substantial agreement; 0.81–0.99 is almost perfect agreement.

- Measure score reliability results

Reliability for each IPF final measure score

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
Between IPFs σ^2	224.4	224.4	224.4	224.4	224.4	224.4	224.4	224.4	224.4
Within IPF σ^2	0.7320	1.1449	1.3456	0.49	1.9321	0.8649	1.5625	0.5625	1.2769
Reliability	0.99675	0.99492	0.99403	0.99782	0.99146	0.99616	0.99309	0.99750	0.99434

- The developer provides explanation for the data elements and IPFs with low Kappa scores.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- *Is the committee concerned about the low Kappa scores for data elements? Is the developer's interpretation reasonable?*
- *Reliability was based on a sample of 100 admissions for each facility. Is it likely that most facilities will have at least 100 admissions during the reporting period?*

Guidance from the Reliability Algorithm

Submitted specifications are precise, unambiguous and complete (Box 1) → Empirical reliability analysis conducted with measure as specified (Box 2) → Reliability testing was conducted with computed performance measure score (Box 4) → Method was appropriate to assess variability in performance at measured entity level (Box 5) → Moderate level of certainty that measure is reliable (Box 6) →Moderate

The highest possible rating is HIGH.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

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

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

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

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- Face validity of the measure score was obtained by a technical expert panel (TEP) vote at the conclusion of measure development and testing. TEP members reviewed responses to each data element, component scores and different summary methods to arrive at the final score.
- TEP voted on the following statement:
  - *"The performance rating from the Medication Reconciliation measure, as specified, represents an accurate reflection of facility-level completeness of the medication reconciliation process on admission to an IPF."*
  - Note: the following reflects NQF guidance on face validity:
    - *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.

Validity testing results:

Face Validity Results by Agreement Category

Agreement Category	Number of Votes	Percent
Agree	6	86%
Disagree	1	14%
Unable to rate	0	0%

Questions for the Committee:

- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *The question to the TEP relates to whether the measure reflects completeness of medication reconciliation, not whether it is able to differentiate quality care. Does the committee agree that the face validity testing meets NQF criteria?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

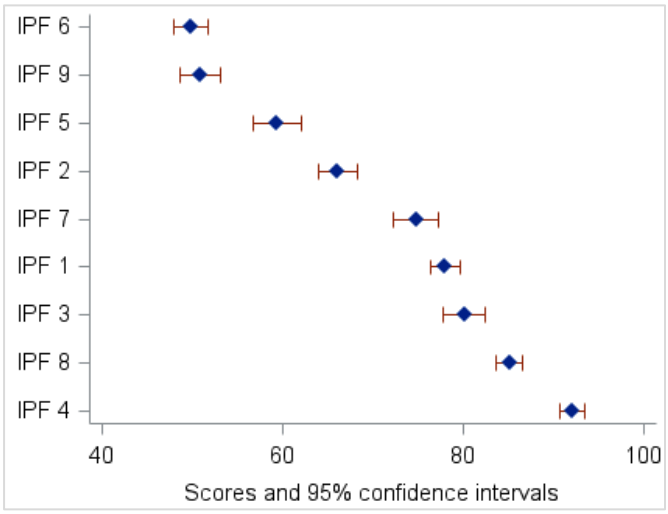
No exclusions.

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

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

- To determine statistically significant differences across the sample of testing facilities, the developer calculated the final scores and 95% confidence intervals for each facility.
- For clinically meaningful differences, the developer reviewed the results of the overall facility level measure scores and the scores of the individual components with their expert workgroup and technical expert panel.

Facility Measure Scores with 95% Confidence Intervals



Question for the Committee:

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

2b6. Comparability of data sources/methods:

Not needed.

2b7. Missing Data

The developer states that this does not apply because the measure score is largely based on the presence of specific data elements in the medical record.

Guidance from the Validity Algorithm

Specifications are consistent with evidence (Box 1)→potential threats assessed (Box 2) →Empirical validity testing not conducted (Box 3)→Face validity assessed on performance measure score (Box 4)→TEP agreed measure score represents completion of medication reconciliation (Box 5) → Moderate

The highest possible rating is MODERATE.

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

2d. Composite measure: Empirical analysis supports construction

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.

- The developer provides a sensitivity analysis for determination of the final score composition (including consideration of different approaches to aggregating scoring elements).
- The developer provides the individual scores by IPF and by data element for Component 1, Component 2, and Component 3.

Final Scores for each component and overall

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Component 1	79.0	32.4	86.8	81.4	37.8	61.5	60.7	79.2	55.3	63.8	32.4, 86.8
Component 2	80.7	89.7	96.1	96.0	76.9	74.0	85.7	76.7	74.2	83.3	74.0, 96.1
Component 3	74.2	76.3	57.5	98.8	63.5	14.0	78.0	99.5	23.3	65.0	14.0, 99.5
Overall Score	78.0	66.1	80.1	92.1	59.4	49.8	74.8	85.1	50.9	70.7	49.8, 92.1
95% CI	76.3, 79.6	64.0, 68.2	77.8 82.5	90.7, 93.4	56.7, 62.1	48.0, 51.7	72.3, 77.2	83.7, 86.6	48.7, 53.1	N/A	N/A

- The developer also provides Pearson correlation coefficients for the association between the three components across facilities to examine whether the components reflect similar performance deficits.
  - For Component 1 and Component 2: 0.31
  - For Component 1 and Component 3: 0.26
  - For Component 2 and Component 3: 0.49
  - Pearson correlations measure the degree of association between two quantitative variables. For the social sciences, scores of 0.37 or larger are considered to have a "large" correlation effect. (Medium effect is 0.24 – 0.36 and small effect is 0.10 – 0.23.)
  - The developer states this correlation analysis supports their integration into a composite measure.

- The developer states the sample size was too small to test for statistical significance.

- The developer notes that in field testing, component 3 called for resolution of PTA medication discrepancies within 24 hours. They modified the specifications (to resolution within 48 hours) after the field testing. The technical expert panel agreed with this change. They added that this change was supported by a sensitivity analysis.

Questions for the Committee:

- *Do the component measures fit the quality construct?*
- *Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*
- *Is the change to Component 3 (resolution of discrepancies within 48 hours) reasonable?*

Preliminary rating for composite construction: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

Comments:

**The performance measure does not measure quality of care or related elements. There are no exclusion? What about a catatonic or non-verbal patients.
**Sample size is small. Results seem to be reliable.
**Definition of "comprehensive" is unclear and appears can only be scored after completing at least component 2. I would think that the scoring would vary based on the completeness of the info and the correctness of the info for component 1 and 2. Not clear how this is related to outcome.

2a2. Reliability Testing

Comments:

**There is a moderate level of reliability.
**Reliability testing was completed with the elements of each of the components. Agreement ranged from 76.6 to 100%, with an average of 91.3%. Kappa scores within facilities demonstrated a range from fair agreement to substantial. The lower scores are concerning.
**The testing was with an adequate number of admissions. Raises a question whether BH facilities would have high volumes of admission to be a reliable metric? Testing was done on a score level.

2b1. Validity Specifications

Comments:

**Face Validity is not strong.
**The specifications are consistent with the evidence.
**Agree with reviews that validity specifications are consistent.

2b2. Validity Testing

Comments:

** Adequate score.
**Face validity done with an expert panel, but a small panel.
**The validity testing was done by face validity testing. The results demonstrate favored voting on the testing methodology.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias

Comments:

**There are no exclusion? What about a catatonic or non-verbal patients.
**Missing data is not applicable.

**I would think that missing data particularly in component 1 would be a major problem and not necessarily an indicator of the lack of the process being performed but the quality of the process.

2d. Composite Analysis
Comments:
**The quality enhancement has not been demonstrated.
**Component measures fit the quality construct. The change in component 3 is reasonable.

Criterion 3. Feasibility

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

- The developer has not specified the measure for sampling, and therefore requires the measure to be calculated for all patients. The developer does note that a sampling strategy may be implemented in the future.
- The developer notes that a structured chart abstraction tool with operational data definitions was developed in Excel for field testing. Prior to implementation, the measure developer states they will provide a finalized abstraction tool.
- Requires data to be manually abstracted from medical records
- Data elements are not currently collected in structured, computer-readable fields

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are there any questions about how the abstraction tool might be modified and how that might impact feasibility?*
- *Is the lack of specification for sampling too burdensome?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy
Comments:
**No concern but is there a need for 3 data elements
**Would like to see a simpler medication reconciliation form.
**The data elements are ideally are routinely generated. I would imagine that the verification of data's completeness may be in question. Medication verification from pharmacy sources is not always available at the moment of treatment in an acute hospitalization unless there is a daily download of pharmacy data that the facility had access to.

Criterion 4: Usability and Use

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

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details

- The developer has submitted this measure to the Measures Under Consideration (MUC) list for use in the Inpatient Psychiatric Facility Quality Reporting Program (IPRQR).

Improvement results N/A

Unexpected findings (positive or negative) during implementation new measure – none reported.

Potential harms none reported

Vetting of the measure none reported

Feedback:

- In 2016, the Measure Application Partnership (MAP) recommended that this measure be refined and resubmitted prior to rulemaking because it is currently undergoing field testing. MAP agreed that testing results should demonstrate reliability and validity at the facility level in the hospital setting. MAP also discussed the intent of the measure (i.e., timeliness vs. accuracy of medication reconciliation) and chart abstraction burden.

***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?*
- *Would inclusion of this measure in the Inpatient Psychiatric Facility Quality Reporting Program be reasonable? Are there other programs that might benefit from this measure?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences

Comments:
**Not vetted.
**This measure is not currently reported.
**The results if reliable could be used to reinforce the adoption of a process by providers and prevent adverse and ineffective treatment.

Criterion 5: Related and Competing Measures

Related or competing measures

0097 : Medication Reconciliation Post-Discharge
0419 : Documentation of Current Medications in the Medical Record
0553 : Care for Older Adults (COA) – Medication Review
0554 : Medication Reconciliation Post-Discharge (MRP)
2456 : Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient

Harmonization

0097 and 0554 focus on reconciliation in the outpatient office/clinic setting and at the health plan level after discharge.
0419 focuses on documenting a patient's medication list in the outpatient office setting.
0553 focuses on annual review of medication list for older adults in various settings at the health plan level.
2456 focuses on documenting the number of unintentional medication discrepancies

A discussion of harmonization is not needed on these measures.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: Face validity only; measure has not been vetted.

Pre-meeting public and member comments

- NAPHS agrees with the concept that quality care is enhanced when there is a clear understanding of a patients' medication history at the time of admission. We acknowledge the careful work done by HSAG in developing this measure. However, we have serious concerns about how the proposed measure will actually be feasible. As specified, data collection for this measure presents an overwhelming burden for facilities and potentially detracts from the organization's ability to focus on the most important elements of mediation reconciliation. There are 21 data elements specified in the measure. While some of these only need to be collected once for the patient, at least 10 must be collected for EACH medication the patient reports. According to the pilot test data, patients reported being on an average of 4.5 medications. In addition, the measure requires verification through a "health systems source" for each medication (whether the patient is deemed to be a reliable informant or not). At the time of admission of a psychiatric patient (always an emergency--there are no scheduled psychiatric admissions) critical decisions must be made about priorities. Compliance with collection of, on average, 44 data elements may seriously interfere with critical priorities. This does not include the burden of seeking information from other health system sources.

  The proposed algorithms and calculations are extremely complex. During the pilot phase, organizations received significant training and were not responsible for calculating performance rates. As this measure potentially rolls out to more than 1600 facilities (with significant turn-over of abstraction staff), it is hard to project how accurate the data will be--even with the very best efforts of clinicians and abstractors. The measure has three very distinct components with complex elements within each component, yet is reported as one measure. This is not typical of other measures used in the CMS payment system requirements.

  Psychiatric hospitals have a significantly lower rate of utilization of electronic health records than general healthcare (there have not been federal funds appropriated to support EHRs in psychiatric hospitals). Electronic systems help, to some degree, with the collection of data (including health system sources). We strongly recommend that the measure also be e-specified to make it usable for facilities that do have electronic medical records.

Many of our members have told us that attempting to find other sources will require "random" calling of pharmacies, outpatient treatment programs, private providers, etc.. Patient consent is required for these contacts. Timely response is totally out of the control of the organization yet a retail pharmacy that is closed is the only consideration given in the measure. Health system sources can often be outdated or incomplete and lack reference to discontinued or over-the counter medications.

The measure materials reference The Joint Commission National Patient Safety Goal on Medication Reconciliation (NPSG.03.05.01). The requirements for meeting this goal (while not designed to be a measure) are much more simple than the proposed CMS measure yet seem to assist facilities in collecting and using relevant data for purpose of safe medication administration.

Failure to meet the requirements of the IPFQR program results in facilities loosing 2% of their Medicare update. Failure to report on one measure constitutes failure to meet the requirements. We cannot support the promulgation of a measure that is inherently so complex, with so many data points and internal inconsistencies, that it presents great potential for organizations to be unsuccessful. The data is publically reported and tracked by many interests. While the data, at this time, is not used for pay for performance purposes, we expect that it eventually will. Benchmarks will be used to determine reimbursement. Collection of data that is not valid and reliable because of problems with the measure, does not allow establishment of benchmarks that are accurate.

We understand the tension between developing the "ideal" measure from a theoretical perspective and the development of a measure that is feasible in the clinical setting in which it is used. We can significantly advance medication reconciliation without overwhelming the field with requirements that distract from answering the most important questions about what medication the patient is on and what they are prescribed during hospitalization.

The burden of both data collection and reporting and the burden of staff training should be evaluated. The numbers reported in the pilot data represent only data abstraction and do not capture the total requirements of the measure reporting. There is a long list of issues in the specifications that we would like to discuss with you in an effort to streamline the measure. Please do not hesitate to call upon us.

Recommendations:

The measure should undergo major simplification, focusing only on data elements that contribute to the most important aspects of medication reconciliation. This process should also result in major simplification of the algorithms and data reporting requirements.

The measure should not be used for payment purposes until it is NQF endorsed. NQF review should include careful attention to the feedback given to CMS/HSAG.

If a patient is deemed to be a reliable informant, no further data sources should be required.

The measure should be e-specified (in addition to non-e-specified) before it is required for the payment program.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

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

Measure Title: Medication Reconciliation on Admission

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

Date of Submission: 12/16/2016

Instructions

- *Complete 1a.1 and 1a.12 for all measures.*
- *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:

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

Notes

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

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

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

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

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

Outcome

☐ Health outcome: Click here to name the health outcome

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

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

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

☐ Structure: Click here to name the structure

☒ Composite: Medication Reconciliation on Admission

1a.12 LOGIC MODEL Diagram or briefly describe the steps between the healthcare structures and processes (e.g., interventions, or services) and the patient's health outcome(s). The relationships in the diagram should be easily understood by general, non-technical audiences. Indicate the structure, process or outcome being measured.

The logic model establishing the process-outcome link for this measure concept is listed below. The process steps corresponding to the measure concept are shown in bold. Literature supporting this logic model is provided in Section 4b.

Patient is admitted for inpatient care → **Care team obtains information on medications the patient was taking prior to admission (PTA) from patient/caregiver(s) and health system sources within 48 hours of admission → Physician reconciles all medications within 48 hours of admission by indicating whether to continue, modify, or discontinue each medication** → Medication errors during the inpatient stay are reduced → Adverse drug events are prevented

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

1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES- State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process (e.g., intervention, or service).

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐US Preventive Services Task Force Recommendation

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

☐Other

The evidence for this measure includes:

- A systematic review published in 2012 and additional studies identified since the review.
- Standards for Medication Reconciliation put forth in the National Patient Safety Goals by The Joint Commission.
- A consensus statement from the Society of Hospital Medicine.

Other Systematic Review

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	Mueller, S. K., Sponsler, K. C., Kripalani, S., & Schnipper, J. L. (2012). Hospital-based medication reconciliation practices: A systematic review. *Archives of Internal Medicine, 172*(14), 1057-1069. doi: 10.1001/archinternmed.2012.2246
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	A systematic review published in 2012 identified 26 controlled studies related to hospital-based medication reconciliation practices (Mueller, Sponsler, Kripalani, & Schnipper, 2012). This review used the 2007 Institute for Healthcare Improvement definition of medication reconciliation, which is the "process of identifying the most accurate list of all medications a patient is taking…and using this list to provide correct medications for patients anywhere within the health care system." The review concludes that the identified studies "consistently demonstrated a reduction in medication discrepancies (17/17 studies), potential adverse drug events (5/6 studies), and adverse drug events (2/2 studies)…Key aspects of successful interventions included intensive pharmacy staff involvement and targeting the intervention to a 'high-risk' patient population." Of note, the systematic review did not discriminate between medication reconciliation at admission, transfer between hospital units, or discharge. *Institute for Healthcare Improvement. (2007) Medication reconciliation review. Retrieved on May 3, 2012 from http: //www.ihi.org/knowledge/Pages/Tools/ MedicationReconciliation Review .aspx
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	The systematic review did not provide an overall grade for the body of evidence. Of the 26 studies identified, 6 were rated as good quality, 5 as fair, and 15 as poor, using the United States Preventive Services Task Force (USPSTF) criteria.

Provide all other grades and definitions from the evidence grading system	The USPSTF grades the quality of the overall evidence for a service on a 3-point scale (good, fair, poor): • Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes. • Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes. • Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.
Grade assigned to the **recommendation** with definition of the grade	The authors of the systematic review did not make specific recommendations.
Provide all other grades and definitions from the recommendation grading system	The authors of the systematic review did not provide other grades and definitions.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	Twenty-six controlled studies were included in the systematic review. Ten of the studies were randomized controlled trials, three were nonrandomized trials with a concurrent control group, and 13 were pre-post studies. Based on the USPSTF grades, 11 of the 26 studies were graded as good to fair quality.
Estimates of benefit and consistency across studies	The studies in the review "consistently demonstrated a reduction in medication discrepancies (17/17 studies), potential adverse drug events (5/6 studies), and adverse drug events (2/2 studies) but showed an inconsistent reduction in postdischarge health care utilization (improvement in 2 of 8 studies)".
What harms were identified?	This review did not identify any harms from this intervention.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We identified new studies by adapting the literature search strategy (medication reconciliation* and patient admission) from Mueller et al. to retreive additional studies that focused on medication reconciliation on admission and were published from November 2010 to present. Using the search terms in PubMed, 277 studies were identified. We included studies written in English and focused on medication reconciliation on admission in hospitalized adults. We excluded studies that evaluated improvement of already existing medication reconciliation processes. Initial independent assessments of titles for relevance and subsequent examination of abstracts resulted in 13 studies retrieved for full-text review. Of these, ten studies were recognized as relevant references. Seven of the studies in the updated review consistently found reductions in medication discrepancies or lower rates of adverse drug events with medication reconciliation on admission as described below: 1. Andreoli et al. (2014) showed a reduction in medication discrepancies by 35%.

2. van den Bemt, van der Schrieck-de Loos, van der Linden, Theeuwes, and Pol (2013) reported a decrease in the percentage of participants with one or more unintentional medication discrepancies from 62% to 32% (OR = 0.29, 95% CI = 0.23-0.37).
3. Chan et al. (2010) measured medication discrepancy rates before and after performing medication reconciliation on admission with the mean rate of 2.6 (SD 2.6) discrepancies per admission dropping to 1.0 (SD 1.1).
4. Giménez Manzorro et al. (2011) implemented a medication reconciliation process at admission and demonstrated a decrease in the rate of medication discrepancies from 7.24% (95% CI = 6.0-8.5) to 4.18% (95% CI = 3.2-5.1).
5. Hron et al. (2015) observed a 53% decrease (P = 0.02) in medication reconciliation errors after implementing an electronic tool on admission.
6. Zoni et al. (2012) reported unintentional medication discrepancies decreased from 3.5% to 1.8% (p< 0.03) when an electronic tool compared a patient's home medication with those prescribed on admission.
7. Boockvar et al. (2011) estimated a 43% reduction in adverse drug events with medication reconciliation on admission.

Three studies did not find associations between medication reconciliation on admission and hospital length of stay or ED revisits after discharge (Hellström , Zoni, Rodríguez Rieiro, et al., 2012; Lisby, Thomsen, Nielsen, et al, 2010; Mendes, Lombardi, Andrzejevski, et al 2016). Of note, measurement of patient outcomes to evaluate the effectiveness of preventive interventions is complicated by the diversity of mechanisms by which inappropriate drug therapy can cause harm.

One additional study was identified through a targeted search for literature specific to the IPF setting (Boswell et. al., 2015). This study found that updating and standardizing medication reconciliation processes increased the accuracy of medications from 45% to 80%.

Citations:

*Andreoli L., Alexandra J.F., Tesmoingt C., Eerdekens, C., Macrez, A., Papo, T.,…Papy, E. (2014). Medication reconciliation: A prospective study in an internal medicine unit. *Drugs and Aging, 31*(5):387-393.

*Boockvar K.S., Blum S., Kugler A.,Livote, E., Mergenhagen, K., Nebeker, J., …Yeh, J. (2011). Effect of admission medication reconciliation on adverse drug events from admission medication changes. *Archives of Internal Medicine,171*(9):860-861.

*Boswell, J.C., Lee, J., Burghart, S.M., Scholtes, K., Miller, L.N. (2015). Medication reconciliation improvement in a private psychiatric inpatient hospital. *Mental Health Clinician*, *5*(1), 35-39. doi:http://dx.doi.org/10.9740/mhc.2015.01.035

	*Chan A.H., Garratt E., Lawrence B., Turnbull, N., Pratapsingh, P., & Black, P. (2010). Effect of education on the recording of medicines on admission to hospital. *Journal of General Internal Medicine, 25*(6):537-542. *Giménez Manzorro, Á., Zoni, A.C., Rodríguez Rieiro, C., Duran- Garcia, E., Lopez, A., Sanz, C.,…Munoz, A. (2011). Developing a programme for medication reconciliation at the time of admission into hospital. *International Journal of Clinical Pharmacy. 33*(4):603-609. *Hellström, L.M., Höglund, P, Bondesson, A., Peterson, G, & Eriksson, T. (2012). Clinical implementation of systematic medication reconciliation and review as part of the Lund Integrated Medicines Management model-impact on all-cause emergency department revisits. *Journal of Clinical Pharmacies and Therapeutics, 37*(6):686-692. *Hron, J.D., Manzi, S., Dionne, R., Chiang, V., Brostoff, M., Altavilla, S., …Harper, M. (2015). Electronic medication reconciliation and medication errors. *International Journal for Quality in Health Care, 27*(4):314-319. *Lisby, M., Thomsen, A., Nielsen, L.P., Lyhne, N., Breum-Leer, C., Fredburg, U.,…Brock, B. (2010). The effect of systematic medication review in elderly patients admitted to an acute ward of internal medicine. *Basic and Clinical Pharmacology and Toxicoogy, 106*(5):422-427. *Mendes, A.E., Lombardi, N.F., Andrzejevski, V.S., Frandoloso, G., Correr, C., & Carvalho, M. (2016). Medication reconciliation at patient admission: a randomized controlled trial. *Pharmacy Practice,* 14(1):656. *van den Bemt, P.M., van der Schrieck-de Loos, E.M., van der Linden, C., Theeuwes, A., & Pol, A. (2013). Effect of medication reconciliation on unintentional medication discrepancies in acute hospital admissions of elderly adults: a multicenter study. *Journal of the American Geriatric Society, 61*(8):1262-1268. *Zoni, A.C., Durán García, M.E., Jiménez Muñoz, A.B., Perez, R., Martin, P., & Alonso, A. (2012). The impact of medication reconciliation program at admission in an internal medicine department. *European Journal of Internal Medicine*, *23*(8):696-700.

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	The Joint Commission. (2016). *National patient safety goals effective January 1, 2017: Hospital Accreditation Program*. Retrieved on December 13, 2016 from https://www.jointcommission.org/assets/1/6/NPSG_Chapter_HAP_Jan2017.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a	The Joint Commission National Patient Safety Goals for hospitals include the following: Maintain and communicate accurate patient medication information (NPSG.03.06.01). The aspects of this goal that are relevant upon admission to the inpatient setting are stated as:

guideline, summarize the conclusions from the SR.	• "Obtain information on the medications the patient is currently taking when he or she is admitted to the hospital or is seen in an outpatient setting. This information is documented in a list or other format that is useful to those who manage medications. Note 1: Current medications include those taken at scheduled times and those taken on an as-needed basis. See the Glossary for a definition of medications. Note 2: It is often difficult to obtain complete information on current medications from a patient. A good faith effort to obtain this information from the patient and/or other sources will be considered as meeting the intent of the EP [element of performance]." • "Define the types of medication information to be collected in non–24-hour settings and different patient circumstances. Note 1: Examples of non–24-hour settings include the emergency department, primary care, outpatient radiology, ambulatory surgery, and diagnostic settings. Note 2: Examples of medication information that may be collected include name, dose, route, frequency, and purpose." • "Compare the medication information the patient brought to the hospital with the medications ordered for the patient by the hospital in order to identify and resolve discrepancies. Note: Discrepancies include omissions, duplications, contraindications, unclear information, and changes. A qualified individual, identified by the hospital, does the comparison."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	None identified
Provide all other grades and definitions from the evidence grading system	Not applicable
Grade assigned to the **recommendation** with definition of the grade	None identified
Provide all other grades and definitions from the recommendation grading system	Not applicable
Body of evidence: • Quantity – how many studies?	From The Joint Commission: "A panel of widely recognized patient safety experts advise The Joint Commission on the development and updating of NPSGs. This panel, called the Patient Safety Advisory Group, is composed of nurses, physicians, pharmacists, risk managers, clinical engineers and other

• Quality – what type of studies?	professionals who have hands-on experience in addressing patient safety issues in a wide variety of health care settings. The Patient Safety Advisory Group works with Joint Commission staff to identify emerging patient safety issues, and advises The Joint Commission on how to address those issues in NPSGs, Sentinel Event Alerts, standards and survey processes, performance measures, educational materials, and Center for Transforming Healthcare projects. Following a solicitation of input from practitioners, provider organizations, purchasers, consumer groups and other stakeholders, The Joint Commission determines the highest priority patient safety issues and how best to address them. The Joint Commission also determines whether a goal is applicable to a specific accreditation program and, if so, tailors the goal to be program-specific." Citation *The Joint Commission. Topic library item. Facts about the National Patient Safety Goals. Development of the Goals. (December 2, 2015). Retrieved November 23, 2016 from https://www.jointcommission.org/facts_about_the_national_patient_safety_goals/
Estimates of benefit and consistency across studies	Not applicable
What harms were identified?	Not applicable
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Not applicable

---

1a.4 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

Not applicable

1a.4.2 What process was used to identify the evidence?

Not applicable

1a.4.3. Provide the citation(s) for the evidence.

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

NQF_Evidence_Attachment-Med_Rec.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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 PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. (Describe how and from whom their input was obtained.)
IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.
This field was left blank per instructions and is included in 1c.3.

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 (4b) under Usability and Use.*
A total of nine inpatient psychiatric facilities (IPFs) from eight different states (AZ, CA, CO, LA, MD, MI, WI, and WV) were used to perform the field testing of the measure. Both freestanding facilities (FS) and hospital-based units of various sizes and with different types of medical records were included in the testing. The sample included a total of 900 admissions from 1/4/2013 through 8/17/2016.

Characteristics of the entities included in the testing are listed below:
IPF ID // Location // Type // Bed Size // Type of Medical Record
1 // WV // Unit // 70 // EPIC
2 // MI // Unit // 28 // McKesson
3 // AZ // FS // 90 // Paper Records
4 // AZ // FS // 75 // Paper Records
5 // MD // FS // 322 // Allscripts®
6 // CA // Unit // 12 // Cerner
7 // LA // Unit // 38 // EPIC
8 // CO // FS // 24 // Netsmart TIER® CareRecord™
9 // WI // FS // 168 // Cerner

Component Scores

Component 1: Comprehensive prior to admission (PTA) medication information gathering and documentation
IPF ID // Component 1 Score (%)
1 // 79.0
2 // 32.4
3 // 86.8
4 // 81.4
5 // 37.8
6 // 61.5
7 // 60.7
8 // 79.2
9 // 55.3
Avg // 63.8

Component 2: Completeness of critical PTA medication information
IPF ID // Component 2 Score (%)
1 // 80.7
2 // 89.7
3 // 96.2
4 // 96.0
5 // 76.9
6 // 74.0
7 // 85.7
8 // 76.7
9 // 74.2
Avg // 83.3

Component 3: Reconciliation action for each PTA medication (within 24 hours)
IPF ID // Component 3 Score (%)
1 // 74.2
2 // 76.3
3 // 57.5
4 // 98.8
5 // 63.5
6 // 14.0
7 // 78.0
8 // 99.5
9 // 23.3
Avg // 65.0

Overall Scores
The final score is calculated for each facility as the average of the scores for each of the three components. Scores for each facility range from 49.9% to 92.1%.

IPF ID // Facility Score // 95% CI
1 // 78.0 // 76.3, 79.6
2 // 66.1 // 64.0, 68.2
3 // 80.1 // 77.8, 82.5
4 // 92.1 // 90.7, 93.4
5 // 59.4 // 56.7, 62.1
6 // 48.8// 48.0, 51.7
7 // 74.8 // 72.3, 77.2
8 // 85.1 // 83.7, 86.6
9 // 50.9 // 48.7, 53.1

Avg // 70.7 // N/A

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
Refer to 1c.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 (4b) under Usability and Use.*
Please refer to Tables 1.6-A (Age and Gender) and 1.6-B (Race and Ethnicity) in the NQF Measure Testing Form for the demographic information of the testing population. Below are the results by gender, age, race, ethnicity and insurance status.

Gender // No. Pts // No. Drugs // Score
Male // 469 // 1962 // 68.3
Female // 431 // 2265 // 69.1

Age // No. Pts // No. Drugs // Score
0-18 // 185 // 328 // 68.3
19-24 // 85 // 195 // 68.4
25-34 // 145 // 453 // 71.3
35-44 // 108 // 598 // 75.7
45-54 // 117 // 652 // 74.7
55-64 // 84 // 606 // 76.9
>65 // 174 // 1391 // 60.5
Missing // 2 // 4 // 84.2

Race // No. Pts // No. Drugs // Score
White // 697 // 3608 // 69.9
Black // 133 // 330 // 65.5
Other // 70 // 289 // 62.5

Ethnicity // No. Pts // No. Drugs // Score
Hispanic // 65 // 164 // 75.7
Non-Hispanic // 750 // 3488 // 70.8
Unknown Ethnicity // 85 // 575 // 52.2

Insurance // No. Pts // No. Drugs // Score
Medicaid // 323 // 877 // 72.4
Medicare // 203 // 1471 // 66.0
Medicare + Medicaid 95 // 799 // 71.4
Other // 279 // 1080 // 69.4

Relative to White patient admissions, Black patients have a 6% lower score and patients with race designated as Other have a 10% lower score. Relative to Hispanic patient admissions, patients with non-Hispanic ethnicity have a 6 % lower score and patients with unknown ethnicity have a 31% lower score relative to Hispanic patients. We do not note differences in scores among female and male patients, but observed 10% lower scores for patients enrolled in Medicare when compared to patients enrolled in Medicaid, perhaps because of age differences rather than insurance status. Relative to children, middle age adults have about 10% higher scores, but geriatric patients have 10% lower scores. Importantly, due to the small sample size, inferences are limited. Multivariate analysis, which are impractical on the facility level given the sample size (i.e., 9 facilities), would be

needed to understand the relationships between age, insurance status and race/ethnicity, given the very diverse demographic distributions across the nine facilities. We will monitor for disparities in a larger sample if the measure is implemented.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b.4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in 1b.4**
Not applicable, see response to 1b.4.

1c. Composite Quality Construct and Rationale

1c.1. **A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.**

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
    - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient);

1c.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1c.2. **Describe the quality construct, including:**

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

This measure calculates the average completeness of the medication reconciliation process within 48 hours of admission to an inpatient facility. The measure was constructed to assess the three elements of The Joint Commission's National Patient Safety Goal (NPSG.03.06.01) on medication safety that are relevant to the admission process (See Section 1a.3). To align with those goals and to simplify the measure's calculation, the measure assesses three components of the medication reconciliation process:

Component 1: Comprehensive prior to admission (PTA) medication information gathering and documentation
Component 2: Completeness of critical PTA medication information
Component 3: Reconciliation action for each PTA medication

It is important to measure all three components because incomplete or inaccurate prior-to-admission (PTA) medication lists may result in inadequate medication reconciliation actions by the prescriber, the ultimate focus of this construct, which aims to prevent medication errors and adverse drug events. According to a 2015 study by the Agency for Healthcare Research and Quality (AHRQ), more than half of admitted patients' medication lists contain at least one discrepancy and 40% of these identified discrepancies have the potential to cause harm (AHRQ, 2015). These errors in prescription medication histories most commonly occur during the admission process (Cornish, et al., 2005). Component 1 addresses this quality gap by ensuring that facilities consult more than one source to obtain patients' prior-to-admission (PTA) medications and that the information is documented in a dedicated area of the medical record for easy reference by providers. Component 2 also addresses this quality gap by encouraging the collection of information on each PTA medications that is necessary to make a reconciliation action, including ordering the medication.

Once the PTA medication information is collected, it is important for the patients' care provider to use that information to inform clinical decision making and reconcile the PTA medications against admission orders. Lack of medication reconciliation during care transitions such as admission to the hospital is responsible for up to 50% of all medication errors and nearly 20% of adverse drug events (ADEs) in the hospital setting (Aspden, Wolcott, Bootman, & Cronenwett, 2007). Component 3 of the measure addresses this quality gap by requiring that a clinician reviews the PTA medications list within 48 hours of admission and documents whether each medication should be continued, discontinued, or modified.

By evaluating not just that a medication reconciliation has been completed but that the medication reconciliation meets several key criteria necessary to reduce medication errors, this measure has the potential to reduce preventable adverse drug events, which are estimated by the Institute of Medicine (IOM) to affect 1.5 million patients per year in the U.S. (Aspden, Wolcott, Bootman, & Cronenwett, 2007).

Citations:
* Agency for Healthcare Research and Quality. (2015). Medication reconciliation. Rockville, MD: U.S. Department of Health and Human Services. Retrieved from: http://psnet.ahrq.gov/primer.aspx?primerID=1.
* Aspden, P., Wolcott, J., Bootman, J. L., & Cronenwett, L. R. (2007). Preventing medication errors: Quality chasm series. Washington, DC: The National Academies Press.
* Cornish, P. L., Knowles, S. R., Marchesano, R., Tam, V., Shadowitz, S., Juurlink, D. N., & Etchells, E. E. (2005). Unintended medication discrepancies at the time of hospital admission. Archives of Internal Medicine, 165(4), 424-429. doi:10.1001/archinte.165.4.424
* Mueller, S. K., Sponsler, K. C., Kripalani, S., & Schnipper, J. L. (2012). Hospital-based medication reconciliation practices: A systematic review. Archives of Internal Medicine, 172(14), 1057-1069. doi: 10.1001/archinternmed.2012.2246
*The Joint Commission. Topic library item. Facts about the National Patient Safety Goals. Development of the Goals. (December 2, 2015). Retrieved November 23, 2016, from https://www.jointcommission.org/facts_about_the_national_patient_safety_goals/

1c.3. **Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.**
The measure was constructed as a composite because of its scoring methodology that first computes scores for each of its three components on the facility level and then averages these scores into a single score. While each component is necessary to describe a single construct, medication reconciliation, presentation of the measure as composite allows close examination of the various parts that define the final single score. Thus, the measure architecture is based on the following considerations:

Alignment with Existing Best Practices: The three components in the measure align with the three core components of Medication Reconciliation on Admission specified by The Joint Commission NPSG.03.06.01.

Simplification of Measure Scoring: A subset of the data elements abstracted for this measure as defined by the data dictionary are used to calculate facility scores. The items in the subset are referred to as scoring elements. Scoring elements in Component 1 are assessed at the medical record level and scoring elements in Components 2 and 3 are assessed at the medication level. The average number of PTA medications per patient varies dramatically across patients as well as across facilities (based on facility case mix, especially related to age). Aggregation of medication information into a single facility-level score for Components 2 and 3 ensures that the scoring elements in Components 2 and 3 contribute consistently to the final measure scores across facilities regardless of the number of medications that were abstracted per record.

1c.4. **Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.**
As described in 1c.2., the composite measure score is calculated from the facility's performance on each of the three components Refer to Section S.14 for the measure algorithm. A total of 11 scoring elements define the medication reconciliation process in this measure. A complete list of the data elements that make up the scoring elements are in Section S.5. Descriptions of the groupings, aggregation, and weighting strategies are included below:

Component 1: Scoring elements of Component 1 are averaged for each record. There are 4 scoring elements for records with medications on the PTA medication list. There are 5 scoring elements for records without medications on the PTA medication list. These record-level averages are then averaged for each facility to produce the facility-level Component 1 score.

Component 2: The five scoring elements of Component 2 are added across all records and divided by the total number of medications abstracted for that facility to produce the facility-level Component 2 score. These scoring elements are averaged across all medications rather than at the record level to ensure that each data element for each medication contributes equally to the overall score regardless of the number of medications per record.

Component 3: A single scoring element assesses the medication reconciliation action step in Component 3. This scoring element is not grouped with the other medication-level scoring elements in Component 2 because the medication reconciliation action is a critical step in the medication reconciliation process and most directly related to the prevention of adverse drug events (i.e., it is not sufficient to merely collect medication information; it needs to inform clinical decision-making). By creating a separate component for the action step, this part of the medication reconciliation process gets equal weight to Components 1 and 2. The Component 3 score is calculated as the proportion of medications across all records that have an action step documented within 48 hours of admission.

Overall Facility-Level Score: The overall composite score is calculated as an average of the three facility-level component scores. Several different weighting approaches were evaluated for each component (see 2d.2). Testing results showed that changes to the weights of the components can impact facility rankings. Based on review of those results, the TEP recommended weighting the components equally.

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

De.7. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Not available. There is no measure-specific webpage.

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:** Med_Rec_Data_Dictionary.xlsx

S.3.1. **For maintenance of endorsement:** Are there changes to the specifications since the last updates/submission. If yes, update the specifications for S1-2 and S4-22 and explain reasons for the changes in S3.2.

S.3.2. **For maintenance of endorsement,** please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
This measure does not have a traditional numerator. The numerator is a facility-level score of the completeness of the medication reconciliation process within 48 hours of admission. This score is calculated by averaging the scores of the three components of the medication reconciliation process. The components include:

1) Comprehensive prior to admission (PTA) medication information gathering and documentation
2) Completeness of critical PTA medication information
3) Reconciliation action for each PTA medication

S.5. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).
The data elements for each component are listed below. Facilities will indicate how many of the criteria were met for each medical record in their sample and follow the measure calculation logic described under the Calculation Algorithm/Measure Logic S.14 to calculate their facility-level score.

Component 1 (C-1): Comprehensive prior to admission (PTA) medication information gathering and documentation:
a) Designated Medication Reconciliation Form/Area: The medical record contains a designated Medication Reconciliation Form/Area that contains a PTA Medication List.
b) Patient source: At least one patient source was referenced to generate the PTA Medication List or the patient was clinically unable to provide medication information and a patient proxy was not available.
-This criterion is met if the medical record contains documentation that the list of medications was supplied by at least one of the following patient sources:
--Interview of the patient or patient proxy
--Medication container brought in by patient or patient proxy
--Medication list brought by patient or patient proxy
--Patient clinically unable to provide medication information and a patient proxy was not available
--Patient support network including a group home
c) Health System Source: At least one health system source was referenced to generate the PTA Medication List.
-This criterion is met if the medical record contains documentation that the list of medications was supplied by at least one of the following health system sources:
--Electronic prescribing network system (e.g., Allscripts®, Surescripts®) or aggregate pharmacy billing (e.g., claims data using state/federal healthcare programs)
--Nursing home
--Outpatient provider
--Outpatient/retail pharmacy or attempt made to contact within 48 hours of admission
--Prescription Drug Monitoring Program (PDMP)
--Prescription in medical record from a prior encounter
d) Contains All H&P Medications: All medications in the History & Physical (H&P) or equivalent document are listed in the PTA Medication List.
e) Medication Reconciliation Form/Area Reviewed by Prescriber Within 48 hours of Admission: When there are no medications on the PTA Medication List, the Medication Reconciliation Form/Area should be reviewed by a licensed prescriber within 48 hours of admission. When there are medications on the PTA Medication List, continue to Components 2 and 3.

Component 2 (C-2): Completeness of critical PTA medication information:
a) Medication Name: This criterion is met when there is a documented medication name.
b) Medication Dose: This criterion is met when there is a valid documented medication dose.
c) Medication Route: This criterion is met when there is a valid documented medication route.
d) Medication Frequency: This criterion is met when there is a valid documented medication frequency.
e) Last Time Medication Taken: This criterion is met when there is a valid documented time when the PTA medication was last taken by the patient or the patient states "unknown" or "cannot remember".

Component 3 (C-3): Reconciliation action for each PTA medication:
a) Timely Reconciled Action: Documentation of reconciliation action to continue, discontinue, or modify the medication
b) Reconciled Action Date and Time: Reconciliation action documented within 48 hours of admission

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The denominator for the composite measure includes admissions to an inpatient facility from home or a non-acute setting with a length of stay greater than or equal to 48 hours.

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 for the composite measure includes admissions to an inpatient facility from home or a non-acute setting with a length of stay greater than or equal to 48 hours.

Admissions to an inpatient facility from home or a non-acute setting
Rationale: Transfers from another inpatient facility or inpatient unit are not included in the denominator population as the medication reconciliation would have been completed by the facility or unit that initially admitted the patient. Long-term care facilities and emergency departments are considered non-acute settings and, therefore, are included in the denominator population.

Admissions with a length of stay greater than or equal to 48 hours
Rationale: Facilities may not have had the chance to complete the medication reconciliation process for admissions with durations shorter than 48 hours.

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
This measure does not have any denominator exclusions.

S.9. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*
Not applicable because there are no denominator exclusions.

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.)*
This measure is not stratified.

S.11. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in measure testing attachment)
No risk adjustment or risk stratification
If other:

S.12. **Type of score:**

Continuous variable, e.g. average
If other:

S.13. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.14. **Calculation Algorithm/Measure Logic** (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*
The algorithm steps are described below. Refer to the supplemental materials for a diagram of the measure logic and data definitions with abstraction instructions for each data element:

Measure Population:
1. Start processing. Run cases that are included in the Initial Patient Population.
2. Check Length of Stay (automatically calculated in hours as equal to the Discharge Date and Discharge Time minus the Admission Date and Admission Time).
a) If the length of stay is greater than or equal to 48 hours, proceed to Transfer From an Acute Care Hospital or Another IPF.
b) If the length of stay is less than 48 hours, the record will be excluded. Stop processing.
3. Check Transfer From an Acute Care Setting.
a) If the patient was admitted from any other admission source, proceed to Component 1 (C-1).
b) If the patient was transferred from an acute care setting, the record will be excluded. Stop processing.

Component 1 (C-1)
1. Check Designated Medication Reconciliation Form/Area.
a. If the Designated Medication Reconciliation Form/Area equals "Yes," proceed to the Medications on PTA Medication List.
b. If the Designated Medication Reconciliation Form/Area equals "No," the C-1 Score for the record will equal zero percent and will count toward the facility's Overall Score. Proceed to Overall Score.
2. Check Medications on PTA Medication List.
a. If Medications on PTA Medication List equals "Yes," the record will receive 25 percentage points for this data element in C-1 (with medications). Proceed to Patient Source (with medications).
b. If Medications on PTA Medication List equals "No," the record will receive 20 percentage points for this data element in C-1 (without medications). Proceed to Patient Source (without medications).
3. Check Patient Source (with medications).
a. If Patient Source (with medications) equals "Yes," the record will receive 25 additional percentage points for this data element in C-1 (with medications). Proceed to Health System Source (with medications).
b. If Patient Source equals "No," the record will receive zero percentage points for this data element in C-1 (with medications). Proceed to Health System Source (with medications).
4. Check Patient Source (without medications).
a. If Patient Source (without medications) equals "Yes," the record will receive 20 additional percentage points for this data element in C-1 (without medications). Proceed to Health System Source (without medications).
b. If Patient Source (without medications) equals "No," the record will receive zero percentage points for this data element in C-1 (without medications). Proceed to Health System Source (without medications).
5. Check Health System Source (with medications).
a. If Health System Source (with medications) equals "Yes," the record will receive 25 additional percentage points for this data element in C-1 (with medications). Proceed to Contains All H&P Medications (with medications).
b. If Health System Source (with medications) equals "No," the record will receive zero percentage points for this data element in C-1 (with medications). Proceed to Contains All H&P Medications (with medications).
6. Check Health System Source (without medications).
a. If Health System Source (without medications) equals "Yes," the record will receive 20 additional percentage points for this data element in C-1 (without medications). Proceed to Contains All H&P Medications (without medications).
b. If Health System Source (without medications) equals "No," the record will receive zero percentage points for this data element in C-1 (without medications). Proceed to Contains All H&P Medications (without medications).

7. Check Contains All H&P Medications (with medications).
a. If Contains All H&P Medications (with medications) equals “Yes,” the record will receive 25 additional percentage points for this data element in C-1 (with medications). Proceed to Component 2 (C-2).
b. If Contains All H&P Medications (with medications) equals “No,” the record will receive zero percentage points for this data element in C-1 (with medications). Proceed to C-2.
8. Check Contains All H&P Medications (without medications).
a. If Contains All H&P Medications (without medications) equals “Yes,” the record will receive 20 additional percentage points for this data element in C-1 (without medications). Proceed to Medication Reconciliation Form/Area Reviewed by Prescriber Within 48 Hours of Admission.
b. If Contains All H&P Medications (without medications) equals “No,” the record will receive zero percentage points for this data element in C-1 (without medications). Proceed to Medication Reconciliation Form/Area Reviewed by Prescriber Within 48 hours of Admission.
9. Check Medication Reconciliation Form/Area Reviewed by Prescriber Within 48 Hours of Admission.
a. If Medication Reconciliation Form/Area Reviewed by Prescriber Within 48 Hours of Admission equals “Yes,” the record will receive 20 additional percentage points for this data element in C-1 (without medications). Proceed to Overall Score for the facility.
b. If Medication Reconciliation Form/Area Reviewed by Prescriber Within 48 Hours of Admission equals “No,” the record will receive zero percentage points for this data element in C-1 (without medications). Proceed to Overall Score for the facility.

Component 2 (C-2)
10. Check Medication Name.
a. For each medication, if the Medication Name equals “Yes or Documented,” the record will receive 20 percentage points for this data element toward C-2. Proceed to Medication Dose.
b. For each medication, if the Medication Name equals “No or Not Documented,” the record will receive zero percentage points for this data element toward C-2. Proceed to Medication Dose.
11. Check Medication Dose.
a. For each medication, if the Medication Dose equals “Yes or Documented,” the record will receive 20 percentage points for this data element toward C-2. Proceed to Medication Route.
b. For each medication, if the Medication Dose equals “No or Not Documented,” the record will receive zero percentage points for this data element toward C-2. Proceed to Medication Route.
12. Check Medication Route.
a. For each medication, if the Medication Route equals “Yes or Documented,” the record will receive 20 percentage points for this data element toward C-2. Proceed to Medication Frequency.
b. For each medication, if the Medication Route equals “No or Not Documented,” the record will receive zero percentage points for this data element toward C-2. Proceed to Medication Frequency.
13. Check Medication Frequency.
a. For each medication, if the Medication Frequency equals “Yes or Documented,” the record will receive 20 percentage points for this data element toward C-2. Proceed to Last Time Medication Taken.
b. For each medication, if the Medication Frequency equals “No or Not Documented,” the record will receive zero percentage points for this data element toward C-2. Proceed to Last Time Medication Taken.
14. Check Last Time Medication Taken.
a. For each medication, if the Last Time Medication Taken equals “Yes or Documented,” the record will receive 20 percentage points for this data element toward C-2. Proceed to step 18.
b. For each medication, if the Last Time Medication Taken equals “No or Not Documented,” the record will receive zero percentage points for this data element toward C-2. Proceed to step 18.
15. Calculate an overall medication score for each medication listed on the PTA Medication list.
16. Determine the medication score for the record by calculating the sum of all medication scores. Proceed to Component 3 (C-3).

Component 3 (C-3)
17. Check Medication Reconciled Action Within 48 Hours of Admission.
a. For each medication, if Reconciled Action and Medication Reconciled Action Within 48 Hours of Admission equals “Yes,” the record will receive 100 percentage points for this data element toward C-3. Proceed to step 21.

b. For each medication, if Reconciled Action and Medication Reconciled Action Within 48 Hours of Admission equals "No," the record will receive zero percentage points for this data element toward C-3. Proceed to step 21.
18. Calculate an overall medication score for each medication listed on the PTA Medication list.
19. Determine the medication score for the record by calculating the sum of all medication scores. Proceed to Overall Score for the facility.

Overall Score
20. Determine the C-1 Score for the facility by calculating the sum of all C-1 scores divided by the total number of records abstracted for C-1.
21. Determine the C-2 Score for the facility by summing all the record-level medication score sums for C-2 and dividing by the total number of medications abstracted in the facility sample.
22. Determine the C-3 Score for the facility by summing all the record-level medication score sums for C-3 and dividing by the total number of medications abstracted in the facility sample.
23. Determine the Overall Score for the facility by calculating the sum of the scores obtained in steps 20, 21, and 22 dividing by 3.

S.15. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Because the measure requires chart review, it will be based on a sample. The measure was evaluated using a random sample of 100 charts per facility, which was an adequate sample size for measure score reliability. The final sampling strategy will be aligned with the current requirements of the IPFQR program with considerations to minimize the abstraction burden for facilities.

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.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable because this measure is not based on patient-reported outcome data.

S.17. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.18.
Other, Paper 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 is collected.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The data dictionary and measure information form that provide instructions for abstracting the data for the measure are included with this application as an attachment. A structured chart abstraction tool with operational data definitions was developed in Excel for field testing. Prior to implementation, the measure developer will provide a finalized abstraction tool.

S.19. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.20. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health : Inpatient
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.)*
Please refer to section S.14. for specific descriptions of steps to calculate each component and overall score for the measure.

2. Validity – See attached Measure Testing Submission Form
NQF_Testing_Attachment-Med_Rec-636175058942657536.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. (Do not remove prior testing information – include date of new information in red.)

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. (Do not remove prior testing information – include date of new information in red.)

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b2-2b7, 2c)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Composite Measure Title: Medication Reconciliation on Admission
Date of Submission: 12/16/2016
Composite Construction:

☒Two or more individual performance measure scores combined into one score

☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)

Medication reconciliation in hospitals occurs at admission, during transfers across hospital units, and on discharge. This measure focuses on medication reconciliation on admission, which is defined by Joint Commission Patient Safety Goals as (1) obtaining information about all medications the patient is taking on admission, (2) defining the types of medications, (3) identifying and resolving discrepancies with medications ordered on admission (the reconciliation action). The proposed measure has been defined accordingly to capture all three steps of medication reconciliation and thus, measures a single construct. The reason this measure is presented as a composite of individual performance measure scores,is inherent in its scoring methodology. The measure averages the score of three components that are first scored on the level of the inpatient psychiatric facility (IPF) and not the patient level.This scoring methodology is chosen because the number of prior-to-admission (PTA) medications varies across patients (e.g., some patients have

none) and because Components 2 (accurate definition of medications) and 3 (reconciliation action) are only applicable if PTA medications are identified. Individual patient scores would vary greatly if the measure were scored on the patient level. To avoid such variation resulting from differences in the number of medications and not in performance, these components are scored at the facility level, thus normalizing the number of PTA medications and balancing the weight of the 3 components.

Given this scoring methodology and following guidance from NQF, we chose a testing methodology that considers each component as a separate performance measure score and then evaluates validity and reliability of the composite. This testing approach is more complex than would be required for "all-or-none" measures, but allows full examination of the "architecture" of the composite.

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- 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.
- **Sections 1, 2a2, 2b2, 2b3, 2b5, 2b7, and 2c must be completed.**
- **For composites with outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b7) and composites (2c) 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 sociodemographic variables and testing in this form refer to the release notes for version 7.0 of the Measure Testing Attachment and the 2016 Measure Evaluation Criteria and Guidance.

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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and

the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;
OR
there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2c2.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(*if not conducted or results not adequate, justification must be submitted and accepted)*

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

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 different components in the composite, indicate the component after the checkbox. If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Not Applicable

1.3. What are the dates of the data used in testing? 1/4/2013 – 8/17/2016

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

A sample of nine Inpatient Psychiatric Facilities (IPFs) from eight different states (AZ, CA, CO, LA, MD, MI, WI, and WV) was used to perform the field testing of the measure. Both free standing facilities and hospital-based units of various sizes and with different types of medical record systems were included in the testing. Table 1.5 provides a breakdown of the characteristics of the IPFs included in the field testing. Each IPF was asked to abstract information from 100 admissions using one of two sampling approaches: (1) selection of most recent admissions or (2) random selection of admissions.

Table 1.5. Field Testing Hospital Characteristics

IPF ID	Location	Type	Bed Size	Data Source
1	WV	Unit	70	EPIC
2	MI	Unit	28	McKesson
3	AZ	FS	90	Paper Medical Records
4	AZ	FS	75	Paper Medical Records
5	MD	FS	322	Allscripts®
6	CA	Unit	12	Cerner
7	LA	Unit	38	EPIC
8	CO	FS	24	Netsmart TIER® CareRecord™
9	WI	FS	168	Cerner

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)?
(*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Testing included a total of 900 admissions from the nine field testing IPFs. The measure considers adult and pediatric patients and has no restriction on insurance type. The only inclusion criterion for testing consisted of admission from home, outpatient, emergency, or long-term care to the IPF for 24 hours or more. The measure does not include transfers because of different care processes involved in medication reconciliation on admission versus upon transfer. The requirement for admission duration was imposed because the measure required resolution of PTA medication discrepancies with admission orders (i.e. the final reconciliation action by a licensed prescriber) within 24 hours. Of note, the measure specifications were modified after field testing was completed to allow for completion of discrepancy resolution within 48 hours of admission. This modification was made in response to TEP concerns regarding accommodating potential delays in obtaining PTA medication information during off-hours when clinics or pharmacies may be closed or due to variable physician staffing schedules at IPFs. The decision was supported by a sensitivity analysis, which identified an increase in scores for Component 3 if 48 instead of 24 hours were allowed. Because a 24 hour timeframe was evaluated during field testing, the field testing results reflect the more stringent 24-hour requirement. Sensitivity analyses comparing 24 and 48 hour turn-around times are presented in Section 2d1.2.

Tables 1.6-A and 1.6-B show the demographic characteristics of the sample by IPF. IPFs varied notably in the proportion of pediatric and geriatric patients as well as the representation of underrepresented minorities, especially African Americans.

Table 1.6-A Age and Gender of Field Testing Population (in percent)

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
No. Records	100	100	100	100	100	100	100	100	100
0-18	0	2	10	45	40	0	4	34	50
19-24	4	20	10	18	12	0	8	5	8
25-34	12	29	28	9	18	0	36	6	7
35-44	18	12	22	11	10	0	23	7	5
45-54	28	19	17	11	9	2	16	12	3
55-64	18	12	11	3	6	5	11	16	2
>65	20	5	1	3	5	93	2	20	25
Male	50	59	55	43	55	44	68	51	44

Table 1.6-B Race/Ethnicity of Field Testing Population (in percent)

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
White	96	72	89	89	60	87	40	93	71
Black	3	7	5	4	31	1	57	3	22
Asian/ Pacific Islander	0	0	1	2	3	6	0	1	0

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
American Indian/ Alaska Native	0	1	5	4	3	1	0	0	1
Other	1	2	0	1	3	5	3	3	0
Unknown Race	0	18	0	0	0	0	0	0	6
Hispanic	1	2	19	24	1	2	3	7	6
Unknown Ethnicity	0	19	0	1	5	55	1	0	4

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Data Element Reliability

A 20% random sample of patient records of the originally sampled 100 records for each facility was re-abstracted by a second independent abstractor for the data element reliability analysis (inter-rater reliability).

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Not applicable, measure is not risk adjusted or stratified.

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

Note*: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. Describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

Two trained abstractors at each IPF independently completed data ascertainment for all measure elements using a random subset of approximately 20 patient records per facility for a total subsample of 175 patient records. There were a total of 5 cases that could not be used for the inter-rater reliability (IRR) testing because these cases had differing admission dates and/or times and could not be matched to cases reviewed by both abstractors.

Table 2a2.2 Distribution of records available for inter-rater reliability analysis across IPFs

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Total
IRR cases	19	20	18	20	20	20	19	19	20	175

Paired abstractors used a structured medical record abstraction tool developed in Microsoft Excel to independently collect data on data elements used to define the measure population and to calculate the measure score. Inter-rater reliability (IRR) between the two abstractors at each site and for each measure data element was assessed using percent overall agreement and Cohen's Kappa statistic. Cohen's Kappa is a measure of inter-rater agreement that accounts for abstractors' agreement by chance alone. It is standardized on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance (i.e., systematic disagreement between abstractors). A common scale is used to interpret Kappa statistics: 0.01–0.20 is considered slight agreement; 0.21– 0.40 is fair agreement; 0.41–0.60 is moderate agreement; 0.61–0.80 is substantial agreement; 0.81–0.99 is almost perfect agreement.

Inter-rater reliability was assessed for each of the 11 data elements that are used to calculate the measure score (5 for Component 1, 5 for Component 2 and 1 for Component 3). We introduced two additional data

elements for purposes of reliability testing only. First, we assessed agreement between reviewers whether the PTA medication list included any medications. One data element in measure Component 1 is applicable only if no medications are listed on the PTA medication list. In these instances, the element measures whether the PTA medication list was reviewed by a licensed prescriber within 24 hours. Measure specifications were updated after field testing was completed and this time interval was increased to 48 hours. For records with PTA medications, this data element is omitted from Component 1 and replaced with Component 3, which requires a reconciliation action by a licensed prescriber within 24 hours. Thus, agreement between reviewers whether the PTA medication list contained medications was necessary to determine whether the element related to 24 hour review should be scored or not. Second, scores for Component 2 and Component 3 of the measure evaluate presence of six data elements (medication name, route, dose, frequency, last time taken and reconciliation action) for each medication listed on the PTA medication list. Because disagreement about the number of medications on the PTA medication list will automatically result in disagreement on all six data elements (in either finding or not finding a given medication), we created for the purposes of inter-rater reliability assessment, a data element for the number of PTA medications. Then, to assess inter-rater reliability for Components 2 and 3, medications that were only identified by one abstractor, were removed from the analysis.

To calculate Cohen's Kappa, we organized the abstractors' responses to all data element questions into four categories (P_{11}: (1, 1), P_{10}: (1, 0), P_{01}: (0, 1) and P_{00}: (0, 0)) for each facility. For each IPF, overall agreement and Cohen's Kappa were calculated for each of the three measure components (by combining all component-specific data elements) and for the total score (by combining all 13 data elements).

We calculated Cohen's Kappa based on the following formula:

$$\text{Cohen's Kappa} = \frac{P_o - P_e}{1 - P_e}$$

In which P_o is the observed proportion of agreement and P_e is the expected proportion of agreement.

$P_o = P_{11} + P_{00}$

$P_e = (P_{11} + P_{10}) * (P_{11} + P_{01}) + (P_{00} + P_{10}) * (P_{00} + P_{01})$

We also report Kappa as aggregate across facilities, separately for each data element, the three components and the final score using Pooled Kappa to account for different rater pairs for each facility.

$$\text{Pooled Kappa} = \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e}$$

In which $\bar{P}_o$ is the mean of the P_os and $\bar{P}_e$ is the mean of the P_es across the nine IPFs (or across a measure component). The 95% confidence intervals of the pooled kappa is $K \pm 1.96 * SE_k$, in which $S_e = \sqrt{\frac{\bar{P}_o(1-\bar{P}_o)}{n(1-\bar{P}_e^2)}}$, and n is the average number of questions across the nine IPFs.

Inter-rater reliability results are shown in Tables 2a2.3-A and 2a2.3-B. "Agreed" means the two abstractors provided consistent answers to the same data element question.

Performance Measure Score Reliability

We used the following formula to calculate the reliability of the score for each IPF, expressed as the signal-to-noise ratio.

$$Reliability = \frac{\sigma^2_{Between-IPFs}}{\sigma^2_{Between-IPFs} + \sigma^2_{Within-IPFs}}$$

In which $\sigma^2_{Between-IPFs}$ is the variance of scores between IPFs and $\sigma^2_{Within-IPFs}$ is the variance within IPFs. The reliability for each IPF score is shown in Table 2a2.3-C.

2a2.3. What were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Table 2a2.3-A. Percent of Agreement and Cohen's Kappa for measure score elements

Data Elements	All Records/ Medications	Agreed	% Agreement	Cohen's Kappa (pooled)^
Component 1				0.66 (0.49, 0.83)
Designated Medication Reconciliation Form/Area	175	166	94.9%	0.67 (0.02, 1.00)
Patient Source	175	134	76.6%	0.20 (-0.46, 0.85)
Health System Source	175	138	78.9%	0.52 (0.11, 0.94)
PTA Medication List Contains All H&P Medications	175	146	83.4%	0.57 (0.14, 1.00)
At least one medication is on PTA Medication List*	175	168	96.0%	0.88 (0.61, 1.00)
PTA Medication List Reviewed by Prescriber within 24 hours of Admission (for records with 0 medications)	42	39	92.9%	0.22 (-1.00, 1.00)
Component 2				0.71 (0.61, 0.81)
Number of Medications on PTA Medication List*	790	701	88.7%	/#
Medication Name	701	701	100.0%	/#
Medication Route	701	695	99.1%	0.91 (0.66, 1.00)
Medication Dose	701	693	98.9%	0.89 (0.61, 1.00)
Medication Frequency	701	685	97.7%	0.67 (0.25, 1.00)
Last Time Medication Taken	701	633	90.3%	0.59 (0.33, 0.84)
Component 3				0.62 (0.36, 0.88)
Medication Reconciliation Action within 24 hours of Admission	701	631	90.0%	0.62 (0.36, 0.88)
Total Score			91.3%	0.73 (0.66, 0.80)

*Added for purposes of reliability testing; not included in the measure score

cannot be calculated because of data structure (e.g., no disagreement or no variation in one category)

^ For simplicity and computational efficiency, we used the normal distribution formula to establish confidence intervals. The confidence intervals based on standard normal distribution may generate upper limits smaller than -1.00 or greater than 1.00, which were adapted to -1.00 and 1.00, respectively.

Table 2a2.3-A. Cohen's Kappa within facilities

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Pooled Kappa
Total Score	0.87 (0.84, 0.91)	0.71 (0.63, 0.78)	0.42 (0.28, 0.55)	0.70 (0.58, 0.83)	0.42 (0.30, 0.53)	0.83 (0.79, 0.87)	0.55 (0.44, 0.66)	0.92 (0.87, 0.96)	0.99 (0.98, 1.00)	0.73 (0.66, 0.80)

Table 2a2.3-C. Reliability for each IPF final measure score

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9
Between IPFs σ^2	224.4	224.4	224.4	224.4	224.4	224.4	224.4	224.4	224.4
Within IPF σ^2	0.7320	1.1449	1.3456	0.49	1.9321	0.8649	1.5625	0.5625	1.2769
Reliability	0.99675	0.99492	0.99403	0.99782	0.99146	0.99616	0.99309	0.99750	0.99434

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Data Element Reliability

The pooled Cohen's Kappa score for the 13 tested data elements across all nine facilities was 0.73 (95% confidence interval: 0.66, 0.80), indicating substantial agreement. The percent agreement across all data elements was high with an average percentage of agreement of 91.3%. The three data elements with less than 90% agreement and lower kappas

include: *Patient Source*, *Health System Source* and *Contains All H&P Medications*. Note that the lower Kappa for *PTA Medication List within 24 hours* has wide confidence intervals due to small sample size, because it is assessed only for PTA medication lists that did not include any medications.

The relatively lower agreement rate in *Patient Source* and *Health System Source* is likely inherent in current medical record documentation practices, which do not require specification of which sources were used to ascertain PTA medications. Thus, abstractors had to read through admission and progress notes to identify potential sources of PTA medications. We anticipate that IPFs will integrate designated fields or check boxes into their medical record if the measure were implemented, which would simplify and standardize data ascertainment.

Regarding the data element *Contains All H&P Medications* , some inconsistencies in abstractors' assessment that the PTA medication list was inclusive of all medications on the H&P resulted from multiple documents that are considered the History & Physical (H&P). Some IPFs noted that in their facility separate admission notes are recorded by a psychiatrist and a general practitioner. To account for this, our measure abstraction instructions were refined to specify that the IPF must identify their principal admission note prior to abstraction so that only one document is used consistently for comparison against the PTA medication list.

Comparing Kappas across IPFs, the measure achieves moderate (IPF 3, 5 and 7), substantial (IPF 2 and 4) and perfect agreement (IPF 1, 6, 8 and 9). Facility 5 identified several reasons for discrepancies including inconsistent documentation practices including inconsistent forms or provider error (e.g., the reconciliation action was dated after discharge). Inconsistencies in IPF 3 and 7 can be explained in part by differing interpretations of admission time, leading to inconsistencies in particular for Component 3. Instructions to use the time of the admission order have been added to the abstraction tool to alleviate this problem.

Performance Measure Score Reliability

Due to the large number of data points for each facility (inherent in the use of individual PTA medications as the unit of analysis in Components 2 and 3, the variance within IPFs is small.The results indicate that the measure score is highly reliable for all nine IPFs included in the sample. All reliability scores are >0.99.

2b2. VALIDITY TESTING

Note: *Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b2.1. What level of validity testing was conducted?

☒ **Composite 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*)

☐ **Systematic assessment of content validity**

☐ **Validity testing for component measures** *(check all that apply)*

Note: *applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Empirical validity testing of the component measure score(s)**

☐ **Systematic assessment of face validity of component measure score(s) as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Systematic Assessment of Face Validity

Face validity of the measure score was obtained by a TEP vote at the conclusion of measure development and testing. The TEP was provided with the final measure specifications and presented the results of field testing. Because the measure score is an aggregate of three components, each scored based on several subcomponents, detailed review of field testing findings was important to assess (a) whether the final score represents an appropriate balance of the various components, (b) whether each subcomponent captures an element of quality and (c) whether each subcomponent shows variation across IPFs and thus, adequate opportunity for improvement. Therefore, TEP members reviewed responses to each data element, component scores and different summary methods to arrive at the final score.

After review and discussion, HSAG asked the TEP members to indicate whether they agreed, disagreed, or were unable to rate the following face validity statement:

"The performance rating from the Medication Reconciliation measure, as specified, represents an accurate reflection of facility-level completeness of the medication reconciliation process on admission to an IPF."

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Systematic Assessment of Face Validity

Seven of 17 members of the IPF TEP were present for the face validity vote. The distribution of the votes is shown in Table 2b2.3.

Table 2b2.3. Face Validity Results by Agreement Category

Agreement Category	Number of Votes	Percent
Agree	6	86%
Disagree	1	14%
Unable to rate	0	0%

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Systematic Assessment of Face Validity

The face validity vote (6/7, 86%) indicates that the measure is viewed as valid by the TEP, which is representative of key stakeholders and experts from the IPF setting.

2b3. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Not applicable because this measure does not have any exclusions.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)
Not applicable because this measure does not have any exclusions.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)
Not applicable because this measure does not have any exclusions.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used? *(check all that apply)*
☐ **Endorsed (or submitted) as individual performance measures**
☒ **No risk adjustment or stratification**
☐ **Statistical risk model**
☐ **Stratification by risk categories**
☐ **Other,** Click here to enter description

2b4.1.1 If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.
Not applicable because this measure is not risk adjusted.

2b4.2. 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 because this measure is not an outcome or resource use measure.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
Not applicable because this measure is not risk adjusted.

2b4.4a. What were the statistical results of the analyses used to select risk factors?
Not applicable because this measure is not risk adjusted.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Not applicable because this measure is not risk adjusted.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix)

below.
If stratified, skip to 2b4.9
Not applicable because this measure is not risk adjusted.

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable because this measure is not risk adjusted.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable because this measure is not risk adjusted.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable because this measure is not risk adjusted.

2b4.9. Results of Risk Stratification Analysis:
Not applicable because this measure is not stratified.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted*)
Not applicable because this measure is not stratified.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not applicable because this measure is not risk adjusted.

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
Note: *Applies to the composite performance measure.*

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
To determine statistically significant differences across the small sample of testing facilities, we calculated the final scores and 95% confidence intervals for each facility, using the following formula:
$S_{\text{final score}} = (S_{C1} + S_{C2} + S_{C3}) / 3$, in which S_{C1} is the score of Component 1, S_{C2} is the score of Component 2, and S_{C3} is the score of Component 3.

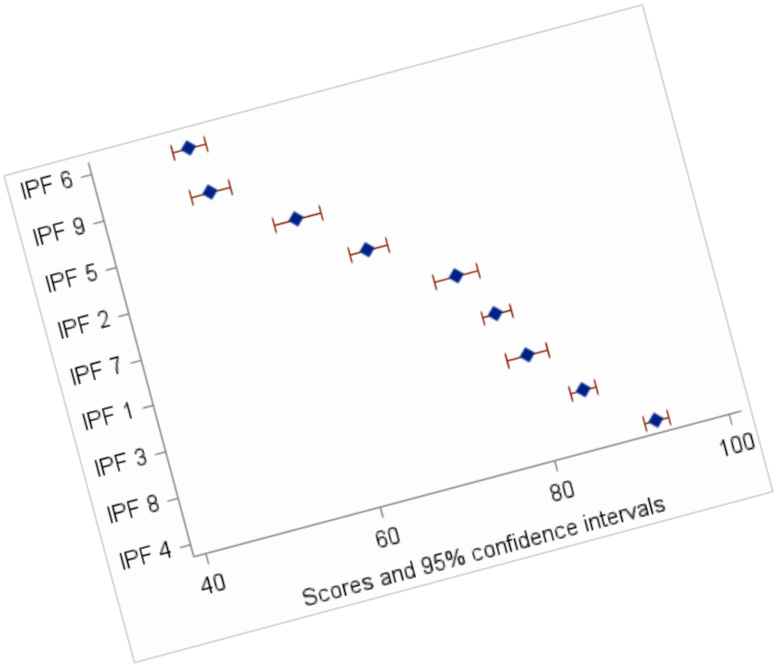
$Se_{\text{final score}} = \sqrt{\frac{1}{9}\frac{S_{C1}\,(100-S_{C1})}{4n_1} + \frac{1}{9}\frac{S_{C2}\,(100-S_{C2})}{5n_2} + \frac{1}{9}\frac{S_{C3}\,(100-S_{C3})}{n_2}}$, in which n_1 is the number of records and n_2 is the number of medications for each IPF. We constructed the Se of the final score based on the assumption that each individual component score represents a proportion of answers with "Yes" to a set of questions. For simplicity, for Component 1, we assume the score is the n1*proportion of answers with "Yes" to four elements. The 95% confidence intervals for the final score is $S_{\text{final score}} \pm 1.96 * Se_{\text{final score}}$. Visual examination of a forest plot depicting measure scores and 95% confidence intervals for each facility can be used to determine whether a given pair of IPFs has statistically significant differences in performance.

For clinically meaningful differences, we reviewed the results of the overall facility level measure scores and the scores of the individual components with our expert workgroup and technical expert panel.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Figure 2b5.2 displays facility scores with 95% confidence intervals sorted by score. Due to small confidence intervals and a wide spread of facility scores, five of the eight adjoining facility pairs have scores with confidence intervals that don't overlap.

Figure 2b5.2. Facility Measure Scores with 95% Confidence Intervals



2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The final IPF facility-level scores indicate substantial variation across facilities with ample room for improvement. Owing to the good precision of the score and the broad range of facility level results, the forest plot illustrates that five of eight adjacent pairs have no overlapping confidence intervals, suggesting statistically significant differences in scores.

As the measure score summarizes the proportion of the medication reconciliation components that were properly completed, the clinical interpretation of differences suggests substantial differences across IPFs in the completeness of information gathering and the reconciliation action.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.* ***If only one set of specifications, this section can be skipped.***

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not***

demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not applicable because only one set of specifications is used.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not applicable because only one set of specifications is used.

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e.*, *what do the results mean and what are the norms for the test conducted?*)
Not applicable because only one set of specifications is used.

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS
Note: *Applies to the overall composite measure.*

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
Note that the measure score is largely based on the presence of specific data elements in the medical record. Thus, missing data issues do not apply.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
Not applicable for the reasons noted in 2b7.1.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
Not applicable for the reasons noted in 2b7.1.

2c. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH
Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.
Common standards such as the Medication Reconciliation standards put forth by The Joint Commission describe this critical care process as a composite of the following steps: (1) obtaining information about all medications the patient is

taking on admission, (2) defining the types of medications, and (3) identifying and resolving discrepancies with medications ordered on admission (the reconciliation action). A measure for this construct should therefore capture the quality and completeness of the gathered PTA medication as well as the timeliness and completeness of reconciliation. Specific challenges in measuring appropriate implementation of these steps include: (1) difficulty to establish a definite list of PTA medications that could serve as a gold standard to evaluate the quality and completeness of the list that was established by the IPF, and (2) challenges to capture from retrospective medical record review whether PTA medications were considered when or soon after admission medications were ordered and whether they were appropriately reconciled. The proposed measure tries to overcome these challenges by defining explicit steps that are expected to result in both, valid PTA medication information and timely reconciliation. As such, each measure element is selected to aid in the operationalization of these processes and therefore a necessary element to measure medication reconciliation. As stated previously, the 11 measure elements are grouped into three components because of different scoring methodologies (and not because they represent separate constructs), but each component has some logical coherence. We introduce each component along with its elements in the following.

Component 1 focuses on the PTA medication gathering process and appropriate integration of that information in provider decision-making. In lieu of a gold standard, accuracy of the PTA medication list is operationalized by requiring that at least one patient and one healthcare system source was considered to gather the information. The component also requires that the PTA medication list has a designated area in the chart for easy reference and that it is inclusive of any PTA medications that the admitted provider may have noted. In order to ensure that the PTA medication list is considered when admission medications are ordered or continued, the measure requires that the list is reviewed by a licensed provider within 48 hours. If the list contains medications, this data element is not considered for scoring and presence of actual reconciliation actions for each listed medication is scored instead (in Component 3). Each element of Component 1 was defined by the measure development work group and reviewed by the TEP and all were found adequate and necessary. It was noted that this measure is the first among those addressing medication reconciliation constructs, that aims to capture the quality of information gathering.

Component 2 formulates a minimum standard of the information needed to identify a unique medication regimen and to facilitate an ordering decision. The first four components *name*, *route*, *dose* and *frequency* are standard elements of every medication order. *Last time taken* is particularly relevant in acute scenarios where standard medication regimen may have been interrupted. Because this element is only ascertainable from the patient, it allows missing information (such as documentation that the patient cannot remember).

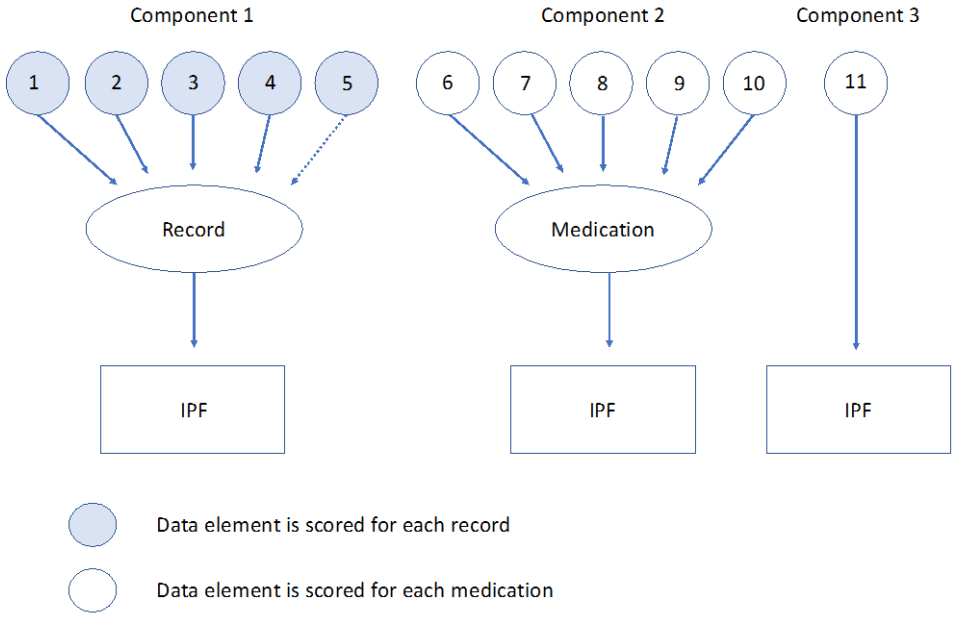
Finally Component 3 captures the reconciliation process by a licensed prescriber within 48 hours of admission. The element requires a decision about continuation, discontinuation or modification for each medication on the PTA medication list. The turn-around time for this item was changed after field testing from 24 to 48 hours because of reports of practical delays in establishing the PTA medication list (e.g., during off hours) and staffing issues (e.g., attending psychiatrist may be present in the IPF for only a confined number of hours). This component is separated from Component 2, even though it is also scored for each medication, because it emphasizes the most important step in the medication reconciliation process.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*
The chosen measure scoring methodology accommodates differing units of analysis for the 11 scoring elements and emphasizes the importance of the actual reconciliation action (Component 3). We chose this methodology together with the measure development work group and the TEP after review of various alternatives and comparison of averages and ranges of the resulting scores across facilities.

Figure 2d1.1 summarizes the measure scoring methodology. For more detailed scoring methodology, please refer to the Measure Information Form in the supplemental materials. Data element scores of Component 1 are averaged for each

record to accommodate differing counts of scoring elements (either 4 scoring elements for records with medications or 5 scoring elements for records without medications on the PTA medication list). These record-level averages are then averaged for each facility (component 1 score). For Component 2, to reduce variation, all medications are pooled across records due to large differences of the number of medications per record. For example, one poorly documented medication on a record with two medications would have a large impact on this record score, while the same poorly documented medication within a record with ten medications would have only a marginal impact. To alleviate this and achieve balance across individual records and ultimately, across IPFs, medications are pooled and each scoring element in component 2 is scored directly on the level of the facility. The scores of the five scoring elements are then averaged (Component 2 score). Finally, Component 3, which consists of one scoring element, is scored for each medication across all records and summarized on the facility level.

Figure 2d1.1 Measure Scoring Methodology



We conducted extensive analyses to arrive at the presented scoring methodology for this measure. For each of the 11 scoring elements we report individual scores (i.e. the proportion of "yes" per facility) and computed averages and ranges across IPFs. For Component 2, we also calculated the average number of medications per record and the proportion of records with no medications, which resulted in the change in scoring methodology that aggregates all medications across records to reduce the impact of the number of medications on measure scores. For Component 3, we conducted a sensitivity analysis that varied the required turn-around time for the reconciliation action from 24 to 48 hours.

Finally, we present Pearson correlation coefficients for the association between the three components across facilities to examine whether the components reflect similar performance deficits.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

The average record scores for Component 1 range from 32.4% to 86.8% (Table 2d1.1-A). The average for each element across facilities ranges from 37.5% for the PTA medication list reviewed by a licensed prescriber within 24 hours data element to 89.7% for the PTA medication list in a designated area data element. Within facilities, record scores range from 0 to 100% in three facilities and, at best, from 50 to 100% in two facilities.

Table 2d1.1-A. Average IPF scores in percent for Component 1

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Designated area	100.0	70.0	100.0	100.0	71.0	89.0	100.0	100.0	77.0	89.7	70.0-100.0
Patient source	89.0	11.0	100.0	97.0	14.0	51.0	80.0	100.0	73.0	68.3	11.0-100.0

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Health system source	84.0	11.0	46.0	38.0	40.0	49.0	35.0	53.0	19.0	41.7	11.0-84.0
PTA med list ⊇ H&P	43.0	37.0	95.0	92.0	22.0	60.0	45.0	61.0	55.0	56.7	22.0-95.0
PTA med list reviewed within 24h of admission (for # meds =0)	33.3	6.1	92.9	69.0	29.8	5.0	0.0	95.0	6.3	37.5	0.0-95.0
# of records with 0 meds	3	33	42	29	47	20	39	20	32	29.4	3, 47
Range of record scores	20, 100	0, 100	50, 100	40, 100	40, 100	0, 100	20, 100	50, 100	0, 100	N/A	N/A
Component 1 score	**79**	**32.4**	**86.8**	**81.4**	**37.8**	**61.5**	**60.7**	**79.2**	**55.3**	**63.8**	**32.4-86.8**

The total number of medications per facility varies across facilities with an average of 2.3 to 9.1 medications per medical record (Table 2d1.1-B). Across facilities, the average completeness is above 90% for all scoring elements with the exception of "last time taken," at 36.4%. The final Component 2 scores range from 74.0% to 96.2% across facilities.

Table 2d1.1-B. Average IPF scores in percent for Component 2

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Total # of meds	913	417	247	320	233	829	241	419	408	447	233, 913
Total # of medical records	100	100	100	100	100	100	100	100	100	100	100, 100
# meds per record	9.1	4.2	2.5	3.2	2.3	8.3	2.4	4.2	4.1	4.5	2.3, 9.1
% records with meds	97	67	58	71	53	80	61	80	68	70.6	53, 97
% Name	99.8	100	100	100	95.7	98.2	99.2	99.3	100	99.1	95.7, 100
% Route	99.6	100	100	99.4	95.7	89.8	98.8	71.6	88.7	93.7	71.6, 100
% Dose	97.4	98.6	95.6	98.4	95.3	91.0	91.7	88.8	91.9	94.3	88.8, 98.6
% Frequency	98.0	99.8	97.6	98.1	95.3	89.6	90.5	78.3	90.4	93.1	78.3, 99.8
% Last time taken	8.7	49.9	87.5	84.1	2.6	1.3	48.1	45.6	0.0	36.4	0, 87.5
Component 2 score	**80.7**	**89.7**	**96.2**	**96.0**	**76.9**	**74.0**	**85.7**	**76.7**	**74.2**	**83.3**	**74.0, 96.2**

The proportion of medications with a reconciled action ranges from 25.9% to 100% across facilities (Table 2d1.1-C). Most facilities denote a time stamp if there was an action, allowing assessment whether the action was completed within 24 hours (as originally required in the measure specifications) or 48 hours (as required in a sensitivity analysis). The requirement for an action within 24 hours shows an average of 65% of all medications across facilities with a range of 14.0% to 99.5%. The requirement for an action within 48 hours shows an average of 71.2% of all medications across facilities with a range of 19.3% to 99.8%.

Table 2d1.1-C. Average IPF scores (in percent) for Component 3

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
%Action	76.1	96.2	100	99.7	91.4	25.9	78.0	99.8	31.4	77.6	25.9, 100
%Action with time	75.9	94.2	94.3	99.7	91.4	22.3	78.0	99.8	31.4	76.3	22.3, 99.8
%Action 24 hours	74.2	76.3	57.5	98.8	63.5	14.0	78.0	99.5	23.3	65.0	14, 99.5
%Action 48 hours	75.9	89.2	69.6	98.8	83.7	19.3	78.0	99.8	26.7	71.2	19.3, 99.8

The final score is calculated for each facility as the average of the scores for each of the three components. Scores for each facility range from 49.8 to 92.1% (Table 2d1.1-D).

Table 2d1.1-D. Final Scores for each component and overall

	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Component 1	79.0	32.4	86.8	81.4	37.8	61.5	60.7	79.2	55.3	63.8	32.4, 86.8
Component 2	80.7	89.7	96.1	96.0	76.9	74.0	85.7	76.7	74.2	83.3	74.0, 96.1

Component 3	74.2	76.3	57.5	98.8	63.5	14.0	78.0	99.5	23.3	65.0	14.0, 99.5
Overall Score	78.0	66.1	80.1	92.1	59.4	49.8	74.8	85.1	50.9	70.7	49.8, 92.1
95% CI	76.3, 79.6	64.0, 68.2	77.8 82.5	90.7, 93.4	56.7, 62.1	48.0, 51.7	72.3, 77.2	83.7, 86.6	48.7, 53.1	N/A	N/A

The Pearson Correlation Coefficients for Component 1 and Component 2, component 1 and Component 3, and Component 2 and Component 3 are 0.31, 0.26 and 0.49, indicating positive correlations among the three measure components across the nine IPFs. Sample size was too small to test for statistical significance.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

The range of scores for each of the five scoring elements in Component 1 shows sufficient variability to discern differences in performance across the nine facilities. Indeed, even within facilities, records show great variability suggesting that processes in each facility are not completely consistent and deserve greater quality assurance.

We observed IPFs performed relatively well on Component 2 in documenting medication name, route, dose and frequency, but variation across facilities ranged by 10 percentage points or more for all elements except the medication name. Documentation of last time taken is generally poor, offering opportunities for improvement for all facilities. We note that improvements in this component will greatly depend on comprehensive efforts in information gathering including both the patient (to ascertain medication taking behavior) and health system sources (to obtain complete medication information).

We emphasize Component 3 as the most critical component because it denotes the ultimate goal of medication reconciliation. The measure scores reflect this importance by averaging results across the three component averages (as opposed to averaging the various elements that are used to score each component). Thus, the proposed scoring methodology gives 33% weight to Component 3, a single item denoting action steps for each medication. Poor scores in this component are driven by two quality issues including general failure to integrate the PTA medication list in the prescriber's work flow and failure to do so in a timely fashion. Two facilities generally failed to have prescribers review and act upon the information on the PTA medication list regardless of turnaround time. Three facilities showed poor scores if a 24-hour timeframe was required but improved in our 48-hour sensitivity analysis, suggesting that work flow and perhaps prescribers' presence at the facility may not always allow faster turnaround.

All three components show variation, suggesting that each adequately contributes to a measure that is aimed to improve quality of the medication reconciliation process. Clinically meaningful differences were observed in the comparison of overall facility level scores and across the range of scores for each of the three components. Components correlate which is supportive of their integration in a composite measure.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To examine the effect of different approaches to aggregate scoring elements we conducted a sensitivity analysis with four different options listed below:

Option 1: Components are ignored, 11 questions are averaged for each facility. This option gives equal weight to the medication reconciliation action (Component 1) and each of the other ten scoring elements summarized in Component 2 and 3.
Option 2: Facility-level components are averaged with more weight on action (¼ – ¼ - ½)
Option 3: Components are scored on the record level and record level scores averaged. This option represents a scoring methodology that most closely reflects the definition of NQF's "all-or none" measures, where several steps of a single care process must be met.
Option 4 (Final score): Facility-level components are averaged (1/3 – 1/3 – 1/3)

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

Table 2d1.-E Comparison of various approaches for final score composition

Scoring	IPF 1	IPF 2	IPF 3	IPF 4	IPF 5	IPF 6	IPF 7	IPF 8	IPF 9	Avg	Range
Option 1	75.2	60.0	88.4	88.6	56.8	58.0	69.7	81.1	56.8	70.5	56.8,88.6
Option 2	77.0	68.6	74.5	93.7	60.4	40.9	75.6	88.7	44.0	68.6	38.2,93.7
Option 3	75.6	50.9	82.4	86.6	40.6	43.4	64.0	82.0	42.8	63.1	40.6,86.6
Option 4 (Final Score)	78.0	66.1	80.1	92.1	59.4	49.8	74.8	85.1	50.9	70.7	49.8,92.1

Option 1: Components are ignored, 11 questions are averaged for each facility
Option 2: Facility-level components are averaged with more weight on action (¼ – ¼ - ½)
Option 3: Components are scored on the record level and record level scores averaged
Option 4 (Final score): Facility-level components are averaged (1/3 – 1/3 – 1/3)

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.e., *what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

Option 1: Slightly lower measure scores observed for the majority of IPFs if the scoring elements are averaged without aggregation in the 3 Components. This option gives lesser weight to the medication reconciliation action (Component 3) which the expert workgroup suggested should be emphasized.
Option 2: Provides larger weight to Component 3 and results in similar scores as the final methodology (Option 4). However, this approach complicates the measure calculation and was considered less feasible.
Option 3: Significant decreases in measure scores observed for eight facilities. Furthermore, the impact of scoring elements with missing information for a single medication will have substantial impact on the record level score for those records with fewer medications and thus distort the overall facility level score.
Option 4 (Final Score): Balanced and most feasible approach that still emphasizes the action step (Component 3)

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

The measure was specified to use manually chart-abstracted data from medical records. This approach was selected for two reasons. First, the setting in which this measure was tested (inpatient psychiatric facilities) primarily used paper medical records at the time of development. Among IPFs that participate in the IPFQR Program, only about 36% attested to using an EHR system for fiscal year 2016 (CMS, 2016).

This approach was also selected because many of the data elements are not currently collected in structured, computer-readable fields. We anticipate that if this measure were to be implemented, some of the data elements could be collected in structured fields. We will further evaluate the feasibility for additional electronic collection if the measure is implemented.

Citation:

* Centers for Medicare & Medicaid Services. Inpatient Psychiatric Facility Quality Measure Data – by Facility. 2016. https://data.medicare.gov/Hospital-Compare/Inpatient-Psychiatric-Facility-Quality-Measure-Dat/q9vs-r7wp. Accessed September 13, 2016.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**

Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
During testing, the average time to abstract each record was 10.6 minutes across all test facilities with a range of 4 minutes per record to 21.5 minutes per record. Facilities that treat patients who are typically on more medications had longer abstraction times due to the time required to abstract information for each medication. We anticipate the average abstraction time will decrease if the measure is implemented as facilities modify their medication reconciliation forms to include some of the data elements in structured fields.

A sampling approach will be selected to minimize the burden of data collection for facilities if the measure is implemented.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees or other requirements to use this measure as specified.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Not in use	

4a.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 because this is a new measure that is not currently in use.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The measure is currently not publicly reported or in use because this is a new measure.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

This measure was submitted by the Centers for Medicare & Medicaid Services (CMS) as a Measure Under Consideration (MUC) for potential inclusion in the Inpatient Psychiatric Quality Reporting (IPFQR) program.

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b. **Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)**

If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

The measure is not in use; therefore, no improvement can be demonstrated at this time. However, information obtained from two of the field testing sites indicated that participating in the field testing of the measure has enabled them to identify opportunities for improvement to their existing medication reconciliation process. Specifically, one site noted that the PTA medications documented in the Emergency Department do not carry over onto the PTA Medication List within the Medication Reconciliation form of their Electronic Health Record, which can result in unreconciled medications. Another site discovered lack of clarity about who (Pharmacy Department or prescribers) was responsible for reconciling the PTA medications. These examples indicate that hospitals can benefit from implementing the measure because it would help reveal opportunities for improvement in their medication reconciliation process. We anticipate that with the implementation of sound medication reconciliation processes improvement in quality of care will be demonstrated over time.

There is empirical evidence that the medication reconciliation process can reduce medication errors and harm across inpatient settings. One study in a Canadian community hospital found that the medication reconciliation process identified and corrected 75% of clinically important medication errors before harm occurred (Vira, Colquhoun, & Etchells, 2006). Another study conducted for acute care hospital patients discharged to long-term care indicated that medication reconciliation reduced medication errors related to ADEs (Boockvar et al., 2006). A study conducted in an IPF found that updating and standardizing its medication reconciliation process resulted in increased accuracy of medications from 45% to 80% based on patient interviews and follow-up with outpatient pharmacies (Boswell, Lee, Burghart, Scholtes, & Miller, 2015). Given that the rate of ADEs is one-third higher in IPFs than in acute care hospitals (Rothschild et al., 2007), improvements to medication reconciliation on admission has the potential to greatly reduce harm to psychiatric patients.

Citations:

*Boockvar, K. S., LaCorte, H. C., Giambanco, V., Fridman, & B., Siu, A. (2006). Medication reconciliation for reducing drug-discrepancy adverse events. The American Journal of Geriatric Pharmacotherapy, 4(3), 236-243. doi:10.1016/j.amjopharm.2006.09.003

*Boswell, J.C., Lee, J., Burghart, S.M., Scholtes, K., & Miller, L.N. (2015). Medication reconciliation improvement in a private psychiatric inpatient hospital. Mental Health Clinician, 5(1), 35-39. doi:http://dx.doi.org/10.9740/mhc.2015.01.035

*Rothschild, J. M., Mann, K., Keohane, C. A., Williams, D. H., Foskett, C., Rosen, S. L., & Bates, D. W. (2007). Medication safety in a psychiatric hospital. General Hospital Psychiatry, 29(2), 156-162. doi:10.1016/j.genhosppsych.2006.12.002

*Vira, T., Colquhoun, M., & Etchells, E. (2006). Reconcilable differences: Correcting medication errors at hospital admission and discharge. Quality and Safety in Health Care, 15(2), 122-126. doi: 10.1136/qshc.2005.015347

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
Not applicable because this measure is not in implementation.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
Not applicable because this measure is not in implementation.

4d1.1. **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.
Individual performance results and assistance with interpretation will be provided to IPFs if the measure is implemented.

4d1.2. **Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.**
Individual performance results and assistance with interpretation will be provided to IPFs if the measure is implemented.

4d2.1. **Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.**
Describe how feedback was obtained.
Feedback from IPFs will be provided during endorsement maintenance if the measure is implemented.

4d2.2. **Summarize the feedback obtained from those being measured.**
Feedback from IPFs will be provided during endorsement maintenance if the measure is implemented.

4d2.3. **Summarize the feedback obtained from other users**
Feedback from other users will be provided during endorsement maintenance if the measure is implemented.

4d.3. **Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
Information on potential measure revisions based on IPF and other user feedback will be provided during endorsement maintenance.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0097 : Medication Reconciliation Post-Discharge
0293 : Medication Information
0419 : Documentation of Current Medications in the Medical Record
0553 : Care for Older Adults (COA) – Medication Review

0554 : Medication Reconciliation Post-Discharge (MRP)
0646 : Reconciled Medication List Received by Discharged Patients (Discharges from an Inpatient Facility to Home/Self Care or Any Other Site of Care)
2456 : Medication Reconciliation: Number of Unintentional Medication Discrepancies per Patient

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?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
The proposed measure is different from the other related measures in several important aspects. First, four of the existing measures are not constructed for the inpatient setting. The proposed measure is specified to address the unique aspects of medication reconciliation in the inpatient setting. Second, the measure focuses on the reconciliation process upon admission to an inpatient facility; whereas, the process measure in the inpatient setting focus on reconciliation at discharge or transfer (NQF 0293 and NQF 0646). Finally, this measure focuses on assessing the quality of the medication reconciliation process versus simply documenting that the process was completed or that medication discrepancies were present. Assessing the reconciliation processes will provide facilities with information on the specific aspects of the medication reconciliation process that can be improved to drive quality.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
This measure complements other existing measures because it focuses on the medication reconciliation process during the first 48 hours of admission to an inpatient facility, which is not addressed by any existing measure. Medication reconciliation at admission is important for accurate medication reconciliation at discharge, which is evaluated by two of the existing measures. Medication reconciliation at admission also ensures that efforts to reconcile medications in the outpatient setting are continued at the transition to the inpatient setting.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment **Attachment:** Supplemental_Materials-Med_Rec.docx

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services, Contracting Officer's Representative (COR)
Co.2 **Point of Contact:** Vinitha, Meyyur, vinitha.meyyur@cms.hhs.gov, 410-786-8819-
Co.3 **Measure Developer if different from Measure Steward:** Health Services Advisory Group, Inc. (HSAG)
Co.4 **Point of Contact:** Megan, Keenan, mkeenan@hsag.com, 616-425-1997-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Technical Expert Panel:
Alisa Busch, MD, MS
Director, Integration of Clinical Measurement & Health Services Research
Chief, Health Services Research Division, Partners Psychiatry and Mental Health
Assistant Professor of Psychiatry and Health Policy, Harvard Medical School

Kathleen Delaney, PhD, PMH-NP, RN
Professor, Rush College of Nursing

Jonathan Delman, PhD, JD, MPH
Assistant Research Professor, Systems and Psychosocial Advances Research Center, University of Massachusetts Medical School

Frank Ghinassi, PhD, ABPP
Vice President, Quality and Performance Measurement, Western Psychiatric Institute and Clinic
Associate Professor in Psychiatry, University of Pittsburg

Eric Goplerud, PhD
Senior Vice President, Director of Public Health Department, NORC at the University of Chicago

Geetha Jayaram, MD
Associate Professor, Schools of Medicine, Health Policy and Management and the Armstrong Institute for Patient Safety, Johns Hopkins University

Charlotte Kauffman, MA, LCPC
Service Systems Coordinator, State of Illinois-Division of Mental Health

Tracy Lenzini, BS
Executive Director, Grand Traverse Health Advocates

Kathleen McCann, RN, PhD
Director of Quality and Regulatory Affairs, National Association of Psychiatric Health Systems

Gayle Olano-Hurt, MPH, CPHQ, PMC
Director Data Management, Outcomes Measurement & Research Administration, Sheppard Pratt Health System

Mark Olfson, MD, MPH
Professor of Psychiatry, Columbia University Medical Center Department of Psychiatry; New York State Psychiatric Institute

Irene Ortiz, MD, MSW
Medical Director, Molina Healthcare of New Mexico

Thomas Penders, MS, MD, DLFAPA
Medical Director, Inpatient Psychiatry, Vident Medical Center
Associate Professor, Brody School of Medicine Department of Psychiatry, East Carolina University

Lucille Schacht, PhD
Senior Director, Performance and Quality Improvement, National Association of State Mental Health Program Directors Research Institute, Inc.

Lisa Shea, MD
Medical Director, Butler Hospital

Thomedi Ventura, MS, MSPH Program Evaluator, Telligen

Elvira Ryan, MBA, BSN, RN
Associate Project Director, Division of Healthcare Quality Evaluation, The Joint Commission

Measure Workgroup:
TEP Members:
Kathleen Delaney, PhD, PMH-NP, RN
Jonathan Delman, PhD, JD, MPH
Irene Ortiz, MD, MSW
Elvira Ryan, MBA, BSN, RN
Lisa Shea, MD

UF Members:
Jordan Daniel Brown, MD
Chief Resident in Adult Psychiatry, Department of Psychiatry, University of Florida College of Medicine

Regina Bussing, MD
Professor and Chair, Department of Psychiatry, University of Florida College of Medicine

Marina Cecchini, MBA
Administrator, UF Health Shands Psychiatric and UF Health Shands Rehab Hospitals

Gigi Lipori, MBA
Chief Data Officer, UF Health and UF Health Sciences Center

Xinyue Liu, PhD
Post-doctoral Fellow

Steve Pittman, PhD
Chief Administrative Officer, Meridian Behavioral Healthcare, Inc.

Ben Staley, PharmD, BCPS
Clinical Specialist, Quality Improvement and Clinical Analytics, Department of Pharmacy UF Health, Shands Hospital

Almut Winterstein, PhD, RPh, FISPE
Professor and Chair, Pharmaceutical Outcomes and Policy, University of Florida College of Medicine

Daniel Zambrano, PharmD
Post-doctoral Fellow

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** Ad.3 **Month and Year of most recent revision:** Ad.4 **What is your frequency for review/update of this measure?** Ad.5 **When is the next scheduled review/update for this measure?**
Ad.6 **Copyright statement:** Not applicable; the measure is in the public domain. Ad.7 **Disclaimers:** None.
Ad.8 **Additional Information/Comments:** None.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3225
Measure Title: Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention
Measure Steward: PCPI Foundation
Brief Description of Measure: Percentage of patients aged 18 years and older who were screened for tobacco use one or more times within 24 months AND who received cessation intervention if identified as a tobacco user
Developer Rationale: This measure is intended to promote adult tobacco screening and tobacco cessation interventions for those who use tobacco products. There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.

Numerator Statement: Patients who were screened for tobacco use at least once within 24 months AND who received tobacco cessation intervention if identified as a tobacco user
Denominator Statement: All patients aged 18 years and older seen for at least two visits or at least one preventive visit during the measurement period
Denominator Exclusions: Documentation of medical reason(s) for not screening for tobacco use (e.g., limited life expectancy, other medical reason)

Measure Type: Process
Data Source: Claims (Only), Claims (Other), Registry
Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Aug 10, 2009 **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 *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.
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:

- The developer states the process/outcome rationale as: *There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke*.
- Clinical practice guidelines from the U.S. Public Health Service (PHS) and recommendations statements from the U.S. Preventive Services Task Force (USPSTF) recommend that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products.

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 following updated USPSTF (2015) statements support the components of this measure:
  - The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and U.S. Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. (“***A” recommendation, “good” or “fair” quality of evidence***)
  - The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. (“***A” recommendation, “good” or “fair” quality of evidence***)
- The developer summarizes the Quality, Quantity and Consistency of evidence to be high across measure components. The developer indicated the review examined the impact of behavioral and pharmacologic interventions on 3 different outcomes:
  - health outcomes including mortality and morbidity
  - tobacco cessation
  - adverse events associated with tobacco cessation interventions

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?*

Guidance from the Evidence Algorithm

Process measure is based on a systematic review (SR) of the evidence and evidence is graded (Box 3) → Summary of QQC of the evidence provided (Box 4) → USPSTF Grade A, with Quantity: high; Quality: high; Consistency: high (Box 5a) → High

The highest possible rating is 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.

-
- Average performance rates, based on data reported to the Physician Quality Reporting System (PQRS)
  - 2011: 81.6%
  - 2012: 84.1%
  - 2013: 89.7%
  - 2014: 88.9% (21.7% of eligible professionals reported on this measure).
- The 2015 mean PQRS ***Claims*** and ***Registry*** performance rates have been provided. For claims, the 4-10th deciles were all performing at 100%, which may suggest little for improvement in those eligible professionals choosing to report the measure. In contrast, for the registry, the 8th-10th deciles performed at 100%.

2015 PQRS Claims Performance Rates

Source	Mean	10th percentile	30th percentile	50th percentile	70th percentile	90th percentile
Claims	96.24%	90.0%	98.3%	100.0%	100.0%	100.0%
Registry	84.36%	51.35%	85.71%	93.25%	98.02%	100.0%

Disparities

- The developer indicates the federal reporting programs in which the measure is utilized have not made disparities data available for analysis and reporting.
- According to published data, disparities exist for counseling (less counseling for Hispanics vs. whites, those who are younger vs. older, and those with worker's compensation or unknown insurance vs. others) and for cessation assistance (higher for those with Medicaid/SCHIP insurance vs. those with private insurance or Medicare and for those who live in a high-poverty area vs. a low-poverty area).

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1.a. Evidence to Support Measure Focus

Comments:

**Evidence remains high and has been updated.

**Evidence is adequate.

**1a. Evidence to Support Measure Focus:
-If measuring a structure, process, or intermediate outcome: How does the evidence relate to the specific structure, process, or intermediate outcome being measured?

There is good evidence for the measure.
-Does it apply directly or is it tangential? How does the structure, process, or intermediate outcome relate to desired outcomes?

It applies directly. If providers intervene then there is a good chance of the patient taking the advice or cessation medication.

For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission?
I am not

If measuring a health outcome or PRO: is the relationship between the measured outcome/PRO and at least one healthcare action (structure, process, intervention, or service) identified AND supported by the stated rationale?
N/A

**Evidence is quit strong with the addition of recent evidence. The quantity , quality and consistency of evidence is high.

1.b. Performance Gap

Comments:

** Yes and disparities based on the published data, I found it interesting that cessation assistance is higher for Medicaid population and those in high poverty areas than private/Medicare and low poverty areas.

**The impact of tobacco use on morbidity and mortality warrants a national measure. The measure shows initial improvement but further improvement needs to occur.

Not enough data to comment on disparities.

**1b. Performance Gap:
-Was performance data on the measure provided?
Yes. For those EPs reporting there is good performance, however, only 21.7 % of eligible EPs actually report on the measure in PQRS.

-How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure?
Many more providers need to adopt the measure

Disparities:
-Was data on the measure by population subgroups provided?
No as that data was not provided to the developer.

-How does it demonstrate disparities in the care?
Other data separate from the Measure Worksheet identify subgroups of the population that receive more or less counseling etc. see pg. 24

**Disparity data is not available at this time. From national statistics there is clear difference in smoking rates by subpopulations beyond sex and age. Differentiation by ethnic group and geo location appears to be apparent.

1.c. Composite

Comments:

**Seems adequate.

**1c. Composite Performance Measure - Quality Construct (if applicable):
- Are the following stated and logical: overall quality construct, component performance measures, and their relationships; rationale and distinctive and additive value; and aggregation and weighting rules?

N/A

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Claims and Registry Data

Specifications:

- This measure is specified for the clinician group/practice level of analysis in the following settings: behavioral health outpatient; clinician office/clinic; home health; occupational therapy evaluation, speech and hearing evaluation, ophthalmological services visit. A higher score indicates better quality.
- The measurement period is a 24-month period.
- For the numerator, "tobacco use" includes any type of tobacco and a "tobacco cessation intervention" includes brief counseling (3 minutes or less) and/or pharmacotherapy. Tobacco use and intervention information are indicated through Category I and II CPT codes.
- The denominator includes patients ages 18 or older, who have had at least 2 visits or 1 preventive care visit during the measurement period. Allowed CPT and HCPCS codes for visits are detailed in the submission.
- Patients can be excluded from the denominator based on "medical reasons" for not screening (e.g., limited life expectancy). This should be coded using a CPT Category II code with modifier 4004F-1P.
- A calculation algorithm is included.
- The measure is not risk adjusted.

Questions for the Committee:

- *Do you have any specific questions on the specifications, codes, definitions, etc.?*
- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment

Maintenance measures – less emphasis if no new testing data provided

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, summarize the reliability testing from the prior review:

- Score-level signal-to-noise analysis was conducted using the beta-binomial method and data from 2011.

Describe any updates to testing:

- Score-level signal-to-noise analysis was conducted using the beta-binomial method and data from 2016. Separate analysis was conducted for claims data and for registry data.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Reliability of the computed measure score was measured as the ratio of signal to noise using a beta-binomial model. The same method was used for both the 2012 evaluation and the current evaluation. This is an appropriate method of assessing score-level reliability.

- A signal-to-noise analysis quantifies the amount of variation in performance that is due to differences between providers (as opposed to differences due to measurement error). Results will vary based on the amount of variation between the providers and the number of patients treated by each provider. The beta-binomial method typically results in a reliability statistic that ranges from 0 to 1 for each provider. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

- Testing for 2012 evaluation
  - The testing sample included 2011 data from 301 physicians and other mid-level providers (e.g., nurse practitioners, midwives, physician assistants) in a large, urban, safety-net network of community health centers in the Midwestern US. The total sample size was 13,312.
  - The developers provided 6 reliability estimates: reliability at the *minimum* number of events and at the *average* number of events, using 3 minimum thresholds for inclusion: n=10 events, n=20 events, and n=30 events.
- Testing for 2017 evaluation
  - Registry testing: The testing sample includes 2015 data reported via the registry option to the PQRS program. This sample included data from 30,033 physicians, of whom 29,949 had all the required data elements and met the minimum number of quality reporting events (n=10). Data for 90.7% of reporting physicians were included in the reliability analysis.
  - Claims testing: The testing sample includes 2015 data reported via the claims option to the PQRS program. This sample included data from 71,445 physicians, of whom 53,326 had all the required data elements and met the minimum number of quality reporting events (n=10). Data for 74.6% of reporting physicians were included in the reliability analysis.
  - The developers provided 2 reliability estimates: reliability at the *minimum* number of events (n=10) and at the *average* number of events.
- NOTE that for testing, clinicians with <10 reporting events in the measurement period were excluded from the analysis. However, the measure specifications do not limit the measure to those with at least 10 reporting events. This means that the reliability estimates from the analysis likely are higher than would be found if all clinicians were included (as typically, reliability increases with sample size).

Results of reliability testing

- Testing for 2012 evaluation

Minimum number of events	Average number of events	Number of eligible providers meeting threshold	Reliability at minimum number of events	Reliability at average number of events
10	76.1	175	0.46	0.86
20	87.8	147	0.61	0.87
30	98.4	126	0.69	0.88

- Testing for 2017 evaluation, based on minimum threshold for inclusion=10 events

Data source	Average number of events	Number of eligible providers meeting threshold	Percent of providers who did not meet threshold	Reliability at minimum number of events	Reliability at average number of events
Registry	312.8	29,949	9.3%	0.78	0.99
Claims	190.5	53,326	25.4%	0.71	0.97

Questions for the Committee:

o *Are there any concerns about the reliability of these measures for providers with very few patients?*

- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm
Submitted specifications are precise, unambiguous and complete (Box 1) → Empirical reliability analysis conducted with measure as specified, except that testing was limited to providers with at least 10 patients (Box 2) → Reliability testing was conducted with computed performance measure score (Box 4) → Method was appropriate to assess variability in performance at measured entity level (Box 5) → Level of certainty that measure is reliable (Box 6) → Moderate

The highest possible rating is HIGH.

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:
- The developer conducted a face validity assessment by a 30-member technical expert panel.

Describe any updates to validity testing:
- The developer conducted another face validity assessment by a 10-member expert panel. Participants included members of the newly-convened PCPI Preventive Care Technical Expert Panel.

SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:
- Face validity for 2012 evaluation
  - After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement: "*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*." Agreement was measured based on a scale from 1 to 5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5= Strongly Agree.
- Face validity for 2017 evaluation
  - Developers used the same method as for the 2012 evaluation.

Validity testing results:

- Face validity for 2012 evaluation
  - N = 17
  - Mean rating = 4.59
  - 16 respondents (94.1%) either agreed or strongly agreed that this measure can accurately distinguish good and poor quality
- Face validity for 2017 evaluation
  - N = 10
  - Mean rating = 3.6
  - 6 respondents (60%) either agreed or strongly agreed that this measure can accurately distinguish good and poor quality

Questions for the Committee:

- *The face validity results declined between the previous and most recent assessment. Is this an indication of diminishing validity of this measure? Or something else?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients can be excluded from the denominator based on "medical reasons" for not screening (e.g., limited life expectancy).
  - Registry (2015 PQRS data): Among the 29,949 physicians with at least 10 quality reporting events:
    - 23,243 exceptions were reported
    - Average number of exceptions per physician= 0.8
    - Overall exception rate=0.2%
  - Claims (2015 PQRS data): Among the 53,326 physicians with at least 10 quality reporting events:
    - 13,762 exceptions were reported
    - Average number of exceptions per physician=0.3
    - Overall exception rate= 0.1%
- The developers note that some have indicated concerns with this kind of "exception" reporting, including potential for gaming by inappropriately excluding patients in order to improve performance results. They cite Doran, et al, 2008 and Kmetik et al., stating that "*research has indicated that levels of exception reporting occur infrequently and are generally valid*".

Questions for the Committee:

- *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)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Data used in the analysis include 2015 PQRS data reported through registry and claims.

Data source	Number of physicians	Mean	Standard Deviation	25th percentile	Median	75th percentile
Registry	29,949	0.84	0.23	0.82	0.93	0.99
Claims	53,326	0.96	0.11	0.97	1.00	1.00

Question for the Committee:
- *Can this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Although the measure can be reported via either claims or registry, no information was provided.

2b7. Missing Data

- No information was provided.

Guidance from the Validity Algorithm
Measure specifications are consistent with evidence provided (Box 1) → Potential threats to validity somewhat assessed (Box 2) → Empirical validity testing NOT conducted (Box 3) → Face validity was systematically assessed (Box 4) → Results indicate substantial agreement that the performance measure score can be used to distinguish quality AND potential threats to validity are likely not a problem (Box 5) → Moderate

The highest possible rating is MODERATE.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.1 & 2b.1 Specifications: Reliability Specifications
Comments:
**a) Can be consistently implemented.

**Specifications are consistent with elements.

**Reliability-Specifications –
Which data elements, if any, are not clearly defined?
They are clear, however, why is this only patients 18 and older? I could not find the reason for that other than where the measure was developed seemed to be in adult settings..

-Which codes with descriptors, if any, are not provided?
Codes seem appropriate

-Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear?
Seems logical

-What concerns do you have about the likelihood that this measure can be consistently implemented?
Will providers really remember to code the "Exceptions as 4004F-1P"?? see pg. 26 Maybe with an HER but if in a registry or other paper process I have doubts as to how many exceptions will actually be captured.

**The exclusions are not clear and rely on clinical judgment of the clinician. The concern would be the ability to work around the numerator if the measure was tied to payment.

2a.2 Reliability Testing
Comments:

**Yes, although I was surprised at the % of providers not meeting threshold using claims data.

**Adequate

**2a2. Reliability - Testing:
-Was reliability tested with an adequate scope (number of entities and patients) to generalize for widespread implementation and with an appropriate method?
Seems this did not address younger people. Might be do to the setting being ones that focus on adults.

Would like to see younger people included as well. AHRQ has a measure that starts at age 12.
https://www.ahrq.gov/sites/default/files/wysiwyg/policymakers/chipra/factsheets/chipra_1516-p003-ef.pdf

-Describe how the results either do or do not demonstrate sufficient reliability.

Seemed to have good reliability testing.
-If a PRO-PM: Was testing conducted at both the data element and score levels?
-If a composite: Was testing conducted at the score level?

**The face validity testing showed variation in results from the two panels with the last panel response implying wider split in opinion. Not clear that the validity testing denotes that this measure has sufficient reliability. Low number of exception in trials.

2b.1 Validity Specifications
Comments:
** Specifications are consistent with evidence.

**2b.1 Validity – Specifications:
-In what ways, if any, are the specifications inconsistent with the evidence?

They are consistent
-If a PRO-PM: In what ways, if any, are the specifications inconsistent with what the target population values and finds meaningful?

**Agree with the reviewer. Evidence is consistent with the evidence.

2b.2 Validity Testing
Comments:
** Used same method as in 2012 but with a decrease in % that agreed or strongly agreed. I don't think it necessarily indicates diminished validy given it was a different and smaller panel.

**No significant problems

**2b2. Validity - Testing:
-Testing:
-Was validity tested with an adequate scope (number of entities and patients) to generalize for widespread implementation and with an appropriate method?

Again – only adults
-Describe how the results either do or do not demonstrate sufficient validity so that conclusions about quality can be made?
Seems to be valid for adults

Why do you agree (or not agree) that the score from this measure as specified is an indicator of quality?
Asking, counseling and providing medication interventions have shown to be effective in getting people to decrease smoking.
-If a PRO-PM: Was testing conducted at both the data element and score levels?

2b.3.-2b7. Testing (Related to Potential Threats to Validity)

Comments:
**I think the exclusion for medical reasons should be reconsidered. I don't think necessarily gaming the system. Is there any rationale for not asking about use of tobacco products and then offering intervention as appropriate.

**No
Hard to comment on disparities

**2b3. Exclusions:
-Are the exclusions consistent with the evidence?
For medical reasons yes. For age I do not think so.

Are any patients or patient groups inappropriately excluded from the measure?
Yes Adolescents

Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?
I am not clear that in ambulatory settings you would have that many patients whre smoking is OK due to a medical condition.

2b4. Risk Adjustment:

If outcome (intermediate, health, or PRO-based) or resource use performance measure:
-Is there a conceptual relationship between potential SDS variables and the measure focus?
Yes. We know certain groups with specific SDS e.g. SMI patients have a higher ragte of smoking that the general population. This measure, however, is not risk adjusted.

How well do SDS variables that were available and analyzed align with the conceptual description provided?
Developer did not provide.

Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)?

Was the risk adjustment (case-mix adjustment) appropriately developed and tested?

Do analyses indicate acceptable results?
-Is an appropriate risk-adjustment strategy included in the measure?

2b5. Meaningful Differences:
-How do analyses indicate this measure identifies meaningful differences about quality?
Measure can help to identify providers that do focus on and assist patients to quit smoking.

2b6. Comparability of performance scores:
-If multiple sets of specifications:
-Do analyses indicate they produce comparable results?
-If risk-adjustment approach includes SDS factors:

Did the developer compare performance scores with and without SDS factors in the risk-adjustment approach?

Did the results support the risk-adjustment approach?

Not risk adjusted

2b7. Missing data/no response:
-Does missing data constitute a threat to the validity of this measure?
I do not think so.

**No data was supplied. Since data sources was from claims and EMR which a consistent source, do not feel that missing data is as big of an issue.

2d. Composite Performance Measure

Comments:
** Would have liked to see relative outcomes of counseling alone, medications alone, and combined counseling and medications clearly listed on cessation rates AND duration of abstinence.

** 2d. Composite Performance Measure - Composite Analysis (if applicable):
-Do analyses demonstrate the component measures fit the quality construct and add value?
-Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

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 developers state that some (not all) data elements are in defined fields in electronic sources. They note that "registry implementation may vary".
- The developer has not identified areas of concern or made modifications based on testing and implementation of the measure.

Questions for the Committee:
- o *Are the required data elements routinely generated and used during care delivery?*
- o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- 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:
**I think this is very feasible and if included in all EHR should be less of a burden.

**3. Feasibility:
-Which of the required data elements are not routinely generated and used during care delivery?
-Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)?
-What are your concerns about how the data collection strategy can be put into operational use?

Will providers really remember to code the "Exceptions as 4004F-1P"?? see pg. 26 Maybe with an HER but if in a registry or other paper process I have doubts as to how many exceptions will actually be captured.

**The data elements have been mostly from electronic sources with some variation in data collection fields from the registries. The electronic submission of data during the practice experience leads to more acurrate alignment with the actions during the encounter with the patient.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No** ☐ **UNCLEAR**

Accountability program details

- This measure is included in the Physician Quality Reporting System, and publicly reported on Physician Compare.
- It is also used in the Million Hearts program.
- Although not mentioned in the submission materials, NQF staff believe this measure is also being used in the following CMS programs:
  - Medicare Shared Savings Program (MSSP)
  - Physician Value-Based Payment Modifier (VBM) [which is being phased out by 12/31/18 and is replaced by MIPS]
  - Physician Feedback/Quality and Resource Use Reports (QRUR) [which is being phased out by 12/31/18 and is replaced by MIPS]

Improvement results

- See performance section under section 1b.
- Although the PQS program has demonstrated increasing performance rates over time which would indicate progress on improvement, it's important to note that the percentage of eligible professional reporting on PQRS measures overall and on this measure, in particular, continues to grow but remains low. In 2014, for example, only 21.7% of eligible professionals reported on the measure. As a result, performance rates may not be nationally representative.

Unexpected findings (positive or negative) during implementation

None reported.

Potential harms None indicated.

Vetting of the measure

- The developer does not discuss any provision of the results and data to those being measured.
- The developer obtains feedback from implementers and those being assessed through a variety of mechanisms.
- In response to feedback from implementers or others, some modifications to the measure specifications and guidance were made prior to finalizing the measure.

Feedback:

- The developers have a process to accept and respond to implementer comments and questions. They report that most questions are for clarification regarding what does, or does not, meet the measure. However, they also note that "*More recently, many implementers wanted to understand how the measure addresses electronic nicotine delivery systems (ENDS*)".

***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

4. Usability and Use:

Comments:

**Is very usable. Do the developers believe the percent of eligible providers reporting will increase with inclusion and use of EHR? I do think the use of ENDS needs to be addressed and we may want to engage the developer in a discussion.

**This is eminently feasible.
The 6 tobacco related NQF endorsed measures should be harmonized and reduced to fewer measures. Suggest we consider recommending 1 hospital level, plan level, adult outpatient level and 1 adolescent level of screening and intervention.
Given the fact that patients with mental illnesses and/or substance use disorders comprise the one population without any progress in this area, I suggest we consider this subset as a disparity group and compare their results with the general population, the poor (MA population), and racial/ethnic sub- groups.

**4. Usability and Use:
-How is the measure being publicly reported?
-For maintenance measures – which accountability applications is the measure being used for?

MACRA/QPP; NYS Medicaid and other programs
How can the performance results be used to further the goal of high-quality, efficient healthcare?

Could be helpful in identifying providers that need assistance in incorporating the measures into practice.
Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.
I do not see any.

Has the measure been vetted in real-world settings by those being measured or others?

If so, has data, results, and aid in interpretation been provided?

Has feedback been solicited?

Was feedback considered if/when changes were made to the measure?

Yes

I do not understand why this measure previously NQF# 0028 is being given a new NQF #. Providers are already using 0028. If we change the NQF number they will lose their history of measure performance. I am not seeing any good rationale for this and would suggest keeping the 0028 #.

Being cynical the developer does make more $$ via licensing fees to technology vendors that incorporate the measure into their programs and these fees are then passed down to the provider community. Something seems wrong with this picture.

Also – I could not tell why on page 25 nicotine electronic delivery systems are being omitted from the measure.

**Measure is already in PQRS and publically reported and part of public health surveillance. Yes feedback was used in evaluation.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0027: Medical Assistance With Smoking and Tobacco Use Cessation
- 1651: TOB-1 - Tobacco Use Screening
- 1654: TOB-2 - Tobacco Use Treatment Provided or Offered
- 1656: TOB-3- Tobacco Use Treatment Provided or Offered at Discharge
- 2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
- 2803 : Tobacco Use and Help with Quitting Among Adolescents
- 3185: Tobacco Use: Screening and Cessation Intervention (eMeasure)

Harmonization

- Measure #0027 is a health plan measure that assess patient-reported advice and advice to quit smoking and other tobacco use, as well as discussion of discussion of cessation strategies and medications. Data for this measure are obtained from the Health Plan CAHPS survey.
- Measures #1651, #1654, and #1656 are hospital-level measures aimed at offering/providing screening, counseling, and cessation interventions
- Measure 2600 focuses on specific populations (SMI, AOD) at the health plan level
- Measure 2803 looks at screening and cessation interventions in adolescents at the clinician group/practice level.
- Measure #3185 is the eMeasure version of this measure. It appears to be harmonized with #3225 to the extent possible.
- These measures seem to be mostly harmonized in terms of their definitions, but potential for further harmonization on specific measures may be discussed at the in-person meeting.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: The measure is not eligible for Endorsement + because empirical validity testing for the measure score has not been conducted.

Pre-meeting public and member comments

- No comments received.

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0028:3225 NQF Project: Behavioral Health

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:

This measure is intended to promote adult tobacco screening and tobacco cessation interventions for those who use tobacco products. There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

The measure focuses on routine tobacco screening for all adults and tobacco cessation interventions for those who use tobacco products. Tobacco use includes use of any type of tobacco.

Clinical practice guidelines from the U.S. Public Health Service (PHS) and recommendations statements from the U.S. Preventive Services Task Force (USPSTF) recommend that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products. The PHS guideline noted that the majority of clinician attention and research in the field has focused on the treatment and assessment of smoking. Nevertheless, they indicated that "[t]he interventions found to be effective in this Guideline have been shown to be effective in a variety of populations. In addition, many of the studies supporting these interventions comprised diverse samples of tobacco users. Therefore, interventions identified as effective in this Guideline are recommended for all individuals who use tobacco, except when medication use is contraindicated or with specific populations in which medication has not been shown to be effective (pregnant women, smokeless tobacco users, light smokers, and adolescents)."

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

In 2015, the USPSTF published an update to its 2009 recommendation on counseling and interventions to prevent tobacco use and tobacco-related disease in adults, including pregnant women. Because there were no plans to update the Public Health Service clinical practice guidelines on treating tobacco use and dependence which formed the basis for the original USPSTF recommendation (2003) and reaffirmation (2009), the Agency for Healthcare Research and Quality (AHRQ) commissioned a new evidence review to assess the benefits and harms of behavioral and pharmacologic interventions for tobacco cessation in adults, including pregnant women. As a result, the 2015 USPSTF updated recommendation is based on the evaluation of evidence summarized in the 2015 review of reviews.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*: Since the measure essentially addresses three components (ie, (1) screening and cessation interventions comprising (2) brief counseling and/or (3) pharmacotherapy), the quantity of studies noted by the guideline are offered as they relate to each of the measure components.

For screening and assessment and its impact on clinical intervention, 9 studies met the selection criteria and were meta-analyzed.

For screening and assessment and its impact on tobacco cessation, 3 studies met the selection criteria and were meta analyzed.

For advice to quit smoking, 7 studies were included in the meta-analysis. For specific information about the intensity of the intervention, namely the efficacy of minimal counseling interventions lasting less than 3 minutes in comparison to low-intensity or high-intensity counseling interventions, 43 studies met the selection criteria for comparison across various lengths.

For combining counseling and medication, 18 studies met selection criteria.

For medication alone, a meta-analysis of 83 studies evaluated the effectiveness and abstinence rates for various medications and medication combinations compared to placebo at 6-months post-quit.

2015 Review of Reviews for the USPSTF

As described above, the evidence review published in 2015 focused on the benefits and harms of behavioral and pharmacologic interventions for tobacco cessation in adults, including pregnant women. It relied primarily on a review of reviews method and included relevant reviews from January 2009 through August 1, 2014.

For behavioral interventions among adults, 26 systematic reviews were included in the analysis.

For pharmacotherapy interventions among adults, 9 systematic reviews were included in the analysis.

For combined pharmacotherapy and behavioral interventions among adults, 1 systematic review was included in the analysis.

For behavioral interventions among pregnant women, 6 systematic reviews were included in the analysis.

For pharmacotherapy interventions among pregnant women, 6 systematic reviews were included in the analysis.

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):* The quality of the body of evidence supporting each of the PHS guideline recommendations is summarized according to the strength of evidence ratings as "A." "A" evidence is described as "Multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings."

Additionally, the medication meta-analysis included predominantly studies with “self-selected” populations. In addition, in medication studies both experimental and control subjects in the studies typically received substantial counseling. Both of these factors tend to produce higher abstinence rates than typically are observed among self-quitters.

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

2015 Review of Reviews for the USPSTF

The quality of the evidence was rated by 2 independent reviewers using a slightly modified version of the AMSTAR (Assessment of Multiple Systematic Reviews) tool. The reviewers then applied the typical USPSTF quality scores (i.e., good-quality, fair-quality, or poor-quality) as described below:

- Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes.
- Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes.
- Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.

All poor quality studies were excluded from the analysis.

For behavioral interventions among adults, 16 systematic reviews were rated as good quality, 10 were rated as fair quality.

For pharmacotherapy interventions among adults, 5 systematic reviews were rated as good quality, 4 were rated as fair quality.

For combined pharmacotherapy and behavioral interventions among adults, 1 systematic review was rated as good quality.

For behavioral interventions among pregnant women, 3 systematic reviews were rated as good quality, 3 were rated as fair quality.

For pharmacotherapy interventions among pregnant women, 5 systematic reviews were rated as good quality, 1 was rated as fair quality.

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):* The consistency of results across studies is summarized according to the strength of evidence ratings as "A." "A" evidence is described as "Multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings."

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

The magnitude and direction of the effect across studies is summarized below for each relevant component addressed by the PHS guideline.

2015 Review of Reviews for the USPSTF

In general, results across all included reviews were consistent within each population and intervention grouping. Reviews rated as good, by definition, include "consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes." The magnitude and direction of the effect for each population and intervention grouping is summarized below.

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

For screening and assessment, the PHS panel looked at two different outcomes - the impact on clinical intervention and tobacco cessation. They concluded that "having a clinic system in place that identifies smokers increases rates of clinician intervention but does not, by itself, produce significantly higher rates of smoking cessation."

Results of the meta-analysis for advice to quit smoking show that brief physician advice significantly increases long-term smoking abstinence rates.

Results of the meta-analysis regarding the intensity of the counseling intervention revealed that all three session lengths (minimal counseling, low-intensity counseling, and higher intensity counseling) significantly increased abstinence rates over those produced by no-contact conditions.

However, there was a clear trend for abstinence rates to increase across these session lengths, with higher intensity counseling producing the highest rates.

For combining counseling and medication, the results of the meta-analysis indicate that providing counseling in addition to medication significantly enhances treatment outcomes.

For medication alone, the PHS Panel identified seven first-line (FDA-approved) medications (bupropion SR, nicotine gum, nicotine inhaler, nicotine lozenge, nicotine nasal spray, nicotine patch, and varenicline) and two second-line (non-FDA-approved for tobacco use treatment) medications (clonidine and nortriptyline) as being effective for treating smokers. Each has been documented to increase significantly rates of long-term smoking abstinence. These medications should be encouraged except where contraindicated or for specific populations for which there is insufficient evidence of effectiveness (i.e., pregnant women, smokeless tobacco users, light smokers, and adolescents).

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

2015 Review of Reviews for the USPSTF

Where possible, the review examined the impact of behavioral and pharmacologic interventions on 3 different outcomes:

- health outcomes including mortality and morbidity
- tobacco cessation
- adverse events associated with tobacco cessation interventions

For behavioral interventions among adults:

- *Health Outcome:* 1 trial found favorable effects on all-cause and coronary disease mortality and lung cancer incidence and mortality 20 years after an intensive behavioral intervention, although results were not statistically significant.
- *Cessation Outcome:* Health provider advice and counseling, tailored self-help materials, and telephone counseling showed modest but significant increased smoking cessation at ≥6 months relative to control participants (18%–96%). Providing more intense adjunctive behavioral support to smokers receiving pharmacotherapy may increase cessation by 9%–24%. Evidence on the use of mobile phone support, Internet-based interventions, and complementary and alternative therapies was limited and not definitive
- *Adverse Event (AE):* Minor AEs related to ear acupuncture, ear acupressure, and other auriculotherapy have been reported. AEs related to other behavioral or complementary and alternative therapies have not been documented.

For pharmacotherapy interventions among adults:

- *Cessation Outcome:* NRT, bupropion SR, and varenicline improve the chances of smoking cessation. Reviews suggested that NRT might increase smoking abstinence at ≥6 mo by 53%–68%, bupropion SR by 49%–76%, and varenicline by 102%–155%. Absolute cessation differences averaged 7% for NRT, 8.2% for bupropion SR, and 26% for varenicline. There were no significant differences among different NRT products, and relative rates of abstinence were similar across settings. Use of a combination of NRT products increases cessation rates more than the use of a single NRT product. In general, there were no significant differences among different classes of medications in direct comparisons.
- *Adverse Event:* NRT, bupropion SR, and varenicline are not associated with an increased risk for major CV AEs. NRT is associated with a higher rate of any CV AE largely driven by low-risk events, typically tachycardia. There was a marginal, nonsignificant increase in serious AEs in participants receiving bupropion SR but no difference for serious psychiatric AEs. The evidence for the safety of varenicline is still under investigation; 1 review suggested a 36% increased risk for nonfatal serious AEs among those receiving varenicline vs. a control intervention.

For combined pharmacotherapy and behavioral interventions among adults,

- *Cessation Outcome:* Combined pharmacotherapy and behavioral interventions increase cessation rates by 70%–100% compared with no or minimal treatment.

For behavioral interventions among pregnant women:

- *Health Outcome:* Statistically significant benefit of behavioral interventions on mean birthweight, low birthweight, and preterm birth vs. usual care or control.
- *Cessation Outcome:* Pooled estimates of a range of behavioral interventions from 70 studies suggested benefits for validated smoking cessation, with a similar benefit when limited to the most common intervention (counseling). Heterogeneity was moderate for the pooled effect, but there was no evidence of subgroup effects by intervention type, number of intervention components, or outcome ascertainment approach.
- *Adverse Event:* No serious AEs reported.

For pharmacotherapy interventions among pregnant women

- *Health Outcome:* Limited evidence of NRT on perinatal and child health benefits. 3 of 4 NRT trials reported fewer preterm births in the intervention group, but only 1 was statistically less than placebo. 2 trials reported higher birthweight in the NRT group; 2 larger trials found no difference. Follow-up data from the largest NRT trial found a higher rate of "survival with no impairment" at 2 y among children of women assigned to the NRT intervention vs. placebo (73% vs. 65%). No trials of bupropion SR or varenicline among pregnant women.
- *Cessation Outcome* No statistical evidence of NRT efficacy for validated smoking cessation in late pregnancy, but power was limited and all trials were in the direction of benefit (pooled analysis based on 5 placebo-controlled trials). No trials of bupropion SR or varenicline among pregnant women.
- *Adverse Event:* No evidence of perinatal harms from NRT. 1 trial found a higher rate of cesarean section for women assigned to NRT; follow-up from the same trial was reassuring for child health outcomes. No trials of bupropion SR or varenicline among pregnant women.

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded? Yes

2015 Review of Reviews for the USPSTF

Yes

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The PHS guideline is the product of a private-sector panel of experts

("the Panel"), representatives of a consortium of several Federal Government and nonprofit organizations, and staff. The panel membership included: Michael C. Fiore, MD, MPH (Panel Chair); Carlos Roberto Jaén, MD, PhD, FAAFP (Panel Vice Chair); Timothy B. Baker, PhD (Senior Scientist); William C. Bailey, MD, FACP, FCCP; Neal L. Benowitz, MD; Susan J. Curry, PhD; Sally Faith Dorfman, MD, MSHSA; Erika S. Froelicher, PhD, RN, MA, MPH;

Michael G. Goldstein, MD; Cheryl G. Healton, DrPH; Patricia Nez Henderson, MD, MPH; Richard B. Heyman, MD; Howard K. Koh, MD, MPH, FACP; Thomas E. Kottke, MD, MSPH; Harry A. Lando, PhD; Robert E. Mecklenburg, DDS, MPH; Robin J. Mermelstein, PhD; Patricia Dolan Mullen, DrPH; C. Tracy Orleans, PhD; Lawrence Robinson, MD, MPH; Maxine L. Stitzer, PhD; Anthony C. Tommasello, PhD, MS; Louise Villejo, MPH, CHES; Mary Ellen Wewers, PhD, MPH, RN.

The evaluation of conflict for the 2008 Guideline Update comprised a two-stage procedure designed to obtain increasingly detailed and informative data on potential conflicts over the course of the Guideline development process. Of the Panel members listed in this document, 21 of 24 had no significant financial interests as defined by the PHS-based criteria. In addition to these mandatory disclosures regarding compensation, leadership, and ownership, members were asked to disclose any other information that might be disclosed in a professional publication. Three Panel members whose disclosures exceeded the PHS criteria for significant financial interest were recused from Panel deliberations relating to their areas of conflict; one additional Panel member voluntarily recused himself.

2015 Review of Reviews for the USPSTF

At least 2 independent reviewers rated the quality of all included systematic review. Discrepancies were resolved through discussion. Additional information regarding disclosures is included in Section 1C.20 below

1c.11 System Used for Grading the Body of Evidence: Other

2015 Review of Reviews for the USPSTF: USPSTF (described in 1c.6.)

1c.12 If other, identify and describe the grading scale with definitions: Every recommendation made by the PHS Panel bears a strength-of-evidence rating that indicates the quality and quantity of empirical support for the recommendation. Each recommendation and its strength of evidence reflects consensus of the Guideline Panel.

The three strength-of-evidence ratings are described as follows:

A. Multiple well-designed randomized clinical trials, directly relevant to

the recommendation, yielded a consistent pattern of findings.

B. Some evidence from randomized clinical trials supported the recommendation, but the scientific support was not optimal. For instance, few randomized trials existed, the trials that did exist were somewhat inconsistent, or the trials were not directly relevant to the recommendation.

C. Reserved for important clinical situations in which the Panel achieved consensus on the recommendation in the absence of relevant randomized controlled trials.

1c.13 Grade Assigned to the Body of Evidence: PHS: A; USPSTF does not separately grade the body of evidence

1c.14 Summary of Controversy/Contradictory Evidence: No controversy or contradictory evidence reported.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:

PHS Guideline (1):

All patients should be asked if they use tobacco and should have their tobacco use status documented on a regular basis. Evidence has shown that clinic screening systems, such as expanding the vital signs to include tobacco use status or the use of other reminder systems such as chart stickers or computer prompts, significantly increase rates of clinician intervention. (Strength of Evidence = A)

All physicians should strongly advise every patient who smokes to quit because evidence shows that physician advice to quit smoking increases abstinence rates. (Strength of Evidence = A)

Minimal interventions lasting less than 3 minutes increase overall tobacco abstinence rates. Every tobacco user should be offered at least a minimal intervention, whether or not he or she is referred to an intensive intervention. (Strength of Evidence = A)

The combination of counseling and medication is more effective for smoking cessation than either medication or counseling alone. Therefore, whenever feasible and appropriate, both counseling and medication should be provided to patients trying to quit smoking. (Strength of Evidence = A)

Clinicians should encourage all patients attempting to quit to use effective medications for tobacco dependence treatment, except where contraindicated or for specific populations for which there is insufficient evidence of effectiveness (i.e., pregnant women, smokeless tobacco users, light smokers, and adolescents). (Strength of Evidence = A)

USPSTF Recommendation (2):

The USPSTF recommends that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products. This is a grade A recommendation.

USPSTF Recommendation (3):

The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and U.S. Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. (A recommendation)

The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. (A recommendation)

1c.17 Clinical Practice Guideline Citation: 1. Fiore MC, Jaen CR, Baker TB, et al. Treating tobacco use and dependence: 2008 update. Clinical practice guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.

2. U.S. Preventive Services Task Force. Counseling and interventions to prevent tobacco use and tobacco-caused disease in adults and pregnant women: U.S. Preventive Services Task Force reaffirmation recommendation statement. Ann Intern Med 2009 Apr 21;150(8):551-5.

3. Siu AL; U.S. Preventive Services Task Force. Behavioral and Pharmacotherapy Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Women: U.S. Preventive Services Task Force Recommendation Statement. Ann Intern Med. 2015 Oct 20;163(8):622-34. doi: 10.7326/M15-2023. Epub 2015 Sep 22.

1c.18 National Guideline Clearinghouse or other URL: www.surgeongeneral.gov/tobacco; www.uspreventiveservicestaskforce.org/uspstf/uspstbac2.htm

https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/tobacco-use-in-adults-and-pregnant-women-counseling-and-interventions1

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded? Yes

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias: The Members of the U.S. Preventive Services Task Force members at the time the 2009 recommendation was finalized represented an array of health-related disciplines including internal medicine, family medicine, behavioral medicine, pediatrics, obstetrics/gynecology and nursing. The task force membership comprised the following individuals: Ned Calonge, MD, MPH, Chair (Colorado Department of Public Health and Environment, Denver, Colorado); Diana B. Petitti, MD, MPH, Vice-Chair (Arizona State University, Phoenix, Arizona); Thomas G. DeWitt, MD (Children´s Hospital Medical Center, Cincinnati, Ohio); Allen J. Dietrich, MD (Dartmouth Medical School, Hanover, New Hampshire); Kimberly D. Gregory, MD, MPH (Cedars-Sinai Medical Center, Los Angeles, California); David Grossman, MD (Group Health Cooperative, Seattle, Washington); George Isham, MD, MS (HealthPartners Inc., Minneapolis, Minnesota); Michael L. LeFevre, MD, MSPH (University of Missouri School of Medicine, Columbia, Missouri); Rosanne M. Leipzig, MD, PhD (Mount Sinai School of Medicine, New York, New York): Lucy N. Marion, PhD, RN (School of Nursing, Medical College of Georgia, Augusta, Georgia); Bernadette Melnyk, PhD, RN (Arizona State University College of Nursing & Healthcare Innovation, Phoenix, Arizona); Virginia A. Moyer, MD, MPH (Baylor College of Medicine, Houston, Texas); Judith K. Ockene, PhD (University of Massachusetts Medical School, Worcester, Massachusetts); George F. Sawaya, MD (University of California, San Francisco, San Francisco, California); J. Sanford Schwartz, MD (University of Pennsylvania Medical School and the Wharton School, Philadelphia, Pennsylvania); and Timothy Wilt, MD, MPH (University of Minnesota Department of Medicine and Minneapolis Veteran Affairs Medical Center, Minneapolis, Minnesota). Prior to each meeting, Task Force members are asked to disclose any information that may interfere with their abilities to discuss and/or vote on a specific topic. Conflicts may arise, for example, if a member has a financial, business/professional, and/or intellectual interest in areas related to a particular topic. All members are expected to provide full disclosure of their interests related to all topics that will be discussed at each meeting. A committee comprised of AHRQ staff and the USPSTF Chair and Vice Chair review each member´s disclosures and issue a recommendation on the member´s eligibility to participate on a specific topic(s). Each member is notified by AHRQ staff of the recommendation prior to each meeting. Members are free to recuse themselves voluntarily from participation in the processes for specific topics; however, a voluntary recusal does not free a member from the obligation to disclose a conflict.

USPSTF 2015 Recommendation Statement

Members of the USPSTF at the time this recommendation was finalized are Albert L. Siu, MD, MSPH, *Chair* (Mount Sinai School of Medicine, New York, and James J. Peters Veterans Affairs Medical Center, Bronx, New York); Kirsten Bibbins-Domingo, PhD, MD, MAS, *Co-Vice Chair* (University of California, San Francisco, San Francisco, California); David Grossman, MD, MPH, *Co-Vice Chair* (Group Health, Seattle, Washington); Linda Ciofu Baumann, PhD, RN, APRN (University of Wisconsin, Madison, Wisconsin); Karina W. Davidson, PhD, MASc (Columbia University, New York, New York); Mark Ebell, MD, MS (University of Georgia, Athens, Georgia); Francisco A.R. García, MD, MPH (Pima County Department of Health, Tucson, Arizona); Matthew Gillman, MD, SM (Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts); Jessica Herzstein, MD, MPH (Independent Consultant, Washington, DC); Alex R. Kemper, MD, MPH, MS (Duke University, Durham, North Carolina); Alex H. Krist, MD, MPH (Fairfax Family Practice, Fairfax, and Virginia Commonwealth University, Richmond, Virginia); Ann E. Kurth, PhD, RN, MSN, MPH (New York University, New York, New York); Douglas K. Owens, MD, MS (Veterans Affairs Palo Alto Health Care System, Palo Alto, and Stanford University, Stanford, California); William R. Phillips, MD, MPH (University of Washington, Seattle, Washington); Maureen G. Phipps, MD, MPH (Brown University, Providence, Rhode Island); and Michael P. Pignone, MD, MPH (University of North Carolina, Chapel Hill, North Carolina). Former USPSTF member Susan Curry, PhD, also contributed to the development of this recommendation.

The USPSTF requires each member to disclose all information regarding any possible financial and nonfinancial conflicts of interest prior to each meeting for all topics under development or that will be discussed at each meeting. Previous disclosures for continuing topics must also be updated to reflect changes in a member's situation since the form was last completed.
Prior to each meeting or to new member appointment, all disclosures are reviewed by the Task Force Chairs according to the criteria specified in the USPSTF Procedure Manual and determined to be either Level 1, 2, or 3. The Task Force Chairs determine the final action on the member's eligibility to participate on a specific topic based on the nature and significance of the potential conflict.

- Level 1 disclosures include nonfinancial disclosures that would not affect the judgment of a Task Force member. These disclosures do not require any action.
- Level 2 disclosures include financial disclosures of $1,000 or less and nonfinancial disclosures that are relevant to a topic but not anticipated to affect the judgment of the Task Force member for that topic. These disclosures are announced at the Task Force meeting, but do not limit the Task Force member's participation in the topic process.
- Level 3 disclosures include financial disclosures of a larger amount and significant nonfinancial disclosures that may affect the Task Force member's view on the topic. Actions for Level 3 disclosures vary according to the nature of the conflict, and may include preventing the member from serving as lead of a topic or on the workgroup of a topic, preventing the member from serving as a primary spokesperson for a topic, or preventing the member from taking part in all topic activities. As all new Task Force members are reviewed for conflicts prior to joining the Task Force, Level 3 disclosures are extremely rare.

1c.21 System Used for Grading the Strength of Guideline Recommendation: USPSTF

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: PHS does not separately grade the strength of the recommendation; USPSTF Grade A

USPSTF 2015 Recommendation Statement

Grade A

1c.24 Rationale for Using this Guideline Over Others: It is the PCPI policy to use guidelines, which are evidence-based, applicable to physicians and other health-care providers, and developed by a national specialty organization or government agency. In addition, the PCPI has now expanded what is acceptable as the evidence base for measures to include documented quality improvement (QI) initiatives or implementation projects that have demonstrated improvement in quality of care.

Recommendations from the USPSTF *are considered the gold standard for clinical preventive services.* The USPSTF is an independent panel of nonfederal experts in prevention and evidence-based medicine. The Task Force carefully assesses the evidence and makes recommendations about preventive services such as screening tests, counseling services, or preventive medications that are provided in clinical settings, and are intended to prevent disease or improve health outcomes from heart disease, cancer, infectious diseases, and other conditions and events that affect the health of children, adolescents, adults, older adults, and pregnant women.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: High 1c.26 Quality: High1c.27 Consistency: High

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

0028_and_3185_Evidence_MSF5.0_Updated_for_2016_Submission-636162857333556000.doc

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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 PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. (Describe how and from whom their input was obtained.)

IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

This measure is intended to promote adult tobacco screening and tobacco cessation interventions for those who use tobacco products. There is good evidence that tobacco screening and brief cessation intervention (including counseling and/or pharmacotherapy) is successful in helping tobacco users quit. Tobacco users who are able to stop smoking lower their risk for heart disease, lung disease, and stroke.

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 (4b) under Usability and Use.*

2014 Physician Quality Reporting System (PQRS) Experience Report

2014 is the most recent year for which PQRS Experience Report measure data are available. The average performance rates on Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention over the last several years are as follows:

- 2011: 81.6%
- 2012: 84.1%
- 2013: 89.7%
- 2014: 88.9%

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. However, beginning in 2015, the program imposed payment penalties for non-participants based on 2013 performance. For 2014, only 21.7% of eligible professionals reported on the measure. As a result, performance rates may not be nationally representative.

Reference: Center for Medicare and Medicaid Services. 2014 Reporting Experience Including Trends. Available: https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/pqrs/analysisandpayment.html

2015 PQRS Claims Performance Rate:
Mean: 96.24%
Minimum: 0.00%
Maximum: 100.00%

Decile	Result %
1	90.0%

2	95.3%
3	98.3%
4	100.0%
5	100.0%
6	100.0%
7	100.0%
8	100.0%
9	100.0%
10	100.0%

2015 PQRS Registry Performance Rate:
Mean: 84.36%
Minimum: 0.00%
Maximum: 100.00%

Decile	Result %
1	51.35%
2	76.92%
3	85.71%
4	90.16%
5	93.25%
6	95.66%
7	98.02%
8	100.00%
9	100.00%
10	100.00%

Report Title: PQRS Ad Hoc Analysis PQ3783, 2015 PQRS Measure Data for PCPI
Report includes 2015 Part B Claims Data for services rendered between January 1, 2015 and December 31, 2015 and processed through February 2016 TAP.
Report also includes PQRS Final Action Registry data and 2015 PQRS Final Action EHR data.

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.**
A number of studies have documented low rates of tobacco use screening and cessation intervention during primary care and other office/outpatient visits, missing key opportunities for intervention.

A 2012 Morbidity and Mortality Weekly Report (MMWR) summarized data from 2005-2008 National Ambulatory Medical Care Survey (NAMCS) and the National Health Interview Survey (NHIS) to determine progress toward Healthy People 2020 objectives calling for increased screening, cessation counseling and cessation success. The following key findings were reported:

- During the study period, adults aged 18 years and older made an estimated annual average of approximately 771 million outpatient visits (an estimated total of 3.08 billion visits during 2005–2008 combined) to office-based physicians.
- Tobacco use screening occurred during the majority of adult visits to outpatient physician offices (62.7%)
- Of the visits that included tobacco use screening, 17.6% (340 million visits) were made by current tobacco users.
- Among patients who were identified as current tobacco users, only 20.9% received tobacco cessation counseling and 7.6% received tobacco cessation medication
- Patients who visited their primary care physician were more likely to receive tobacco screening (66.6% of visits) than patients who visited a physician who was not their primary care physician (61.6% of visits). Screening also varied by physician specialty. Patients visiting general or family practitioners (66.4%) and obstetricians/gynecologists (69.6%) were more likely to receive screening than patients who visited physicians in other specialties (58.2%), excluding internal medicine, cardiovascular disease, and psychiatry. (1)

Given that hospital outpatient visits account for approximately 1 in 10 outpatient visits, Jamal and colleagues sought to assess the rates of tobacco use screening and cessation assistance offered to US adults during their hospital outpatient clinic visits analyzing data from the 2005–2010 NAMCS. The following key findings were reported:

- During the study period, adults aged 18 years or older made, on average, 71.8 million hospital outpatient visits annually to hospital outpatient physicians or an estimated 431 million visits from 2005 through 2010 combined.
- On average, 45.2 million (63.0%) hospital outpatient visits included tobacco use screening each year.
- Of the visits that included tobacco use screening, 25.7% (11.6 million annual average visits) were made by current tobacco users.
- Among patients who screened positive for current tobacco use, 24.5% (or an estimated 17.1 million visits) received any cessation assistance, including tobacco counseling, a prescription or order for a cessation medication at the visit, or both.
- Patients who made visits to general medicine clinics (67.1%) were more likely to receive tobacco use screening than those who made visits to surgical clinics (55.7%) or clinics with other specialties (45.2%), excluding obstetrics and gynecology (62.8%) and substance abuse clinics (68.3%). (2)

Citations:

1. Jamal A1, Dube SR, Malarcher AM, Shaw L, Engstrom MC; Centers for Disease Control and Prevention (CDC). Tobacco use screening and counseling during physician office visits among adults--National Ambulatory Medical Care Survey and National Health Interview Survey, United States, 2005-2009. MMWR Suppl. 2012 Jun 15;61(2):38-45.
2. Jamal A, Dube SR, King BA. Tobacco Use Screening and Counseling During Hospital Outpatient Visits Among US Adults, 2005–2010. Prev Chronic Dis 2015;12:140529.

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 (4b) 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**

The MMWR noted that rates of tobacco screening and intervention varied by patients' race, age and insurance status. Overall, patients classified as non-Hispanic whites were more likely to receive counseling than Hispanic patients (64.1 versus 57.8%). Among current tobacco users, younger patients (aged 25 to 44 years) reported receiving less counseling (17.9%) than patients aged 45 to 64 years (22.7%). Patients with workers' compensation, and those whose insurance status was unknown were less likely to receive counseling than those with private insurance, self-payers, Medicaid, and Medicare patients.

Similar racial/ethnic disparities were reported for hospital outpatient visits. Tobacco use screening varied by patient's race/ethnicity - visits made by Hispanics (55.4%) were less likely to receive tobacco use screening than those by non-Hispanic whites (65.1%). For tobacco users, cessation assistance was higher for visits made by those with Medicaid/SCHIP (27.6%) than those with private insurance (21.8%) or Medicare (21.4%). Patients living in a high poverty zone were more likely to receive cessation than those living in a low poverty zone. (2)

1. Jamal A1, Dube SR, Malarcher AM, Shaw L, Engstrom MC; Centers for Disease Control and Prevention (CDC). Tobacco use screening and counseling during physician office visits among adults--National Ambulatory Medical Care Survey and National Health Interview Survey, United States, 2005-2009. MMWR Suppl. 2012 Jun 15;61(2):38-45.
2. Jamal A, Dube SR, King BA. Tobacco Use Screening and Counseling During Hospital Outpatient Visits Among US Adults, 2005–2010. Prev Chronic Dis 2015;12:140529.

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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 : 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.)*
The measure specifications are included with this submission. Additional measure details may be found at http://www.thepcpi.org/pcpi/media/PCPI-Maintained-Measures/Preventive-Care-and-Screening-Updated-June-2016.pdf.

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:** NQF0028_CMS138v5_ValueSets_Details.xlsx

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.
Supporting guidelines and coding included in the measure are reviewed on an annual basis. The updated recommendation from USPSTF published in 2015 resulted in updated clinical recommendation statements and guidance regarding the intentional omission of electronic nicotine delivery systems (ENDS) from the measure. Additional limited changes have been incorporated into the technical specifications to adhere to current industry standards while preserving the original measure intent.

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 who were screened for tobacco use at least once within 24 months AND who received tobacco cessation intervention if identified as a tobacco user

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 least once during the 24 month period

Definitions:
Tobacco Use – Includes any type of tobacco
Tobacco Cessation Intervention – Includes brief counseling (3 minutes or less), and/or pharmacotherapy

For Administrative Claims/Registry:
CPT Category II code 4004F: Patient screened for tobacco use AND received tobacco cessation intervention (counseling, pharmacotherapy, or both), if identified as a tobacco user
OR
CPT Category II code 1036F: Current tobacco non-user
OR
CPT Category I code- Smoking and tobacco use cessation counseling
*The following codes are applicable if the patient screened positive for smoking/tobacco use and counseling was provided.
99406: Smoking/tobacco counseling 3-10 minutes
99407: Smoking/tobacco counseling greater than 10 minutes

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients aged 18 years and older seen for at least two visits or at least one preventive visit during the measurement period

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

For Administrative Claims/Registry:
Patient age >= 18 years
AND
At least two visits during the measurement period (CPT):
90791, 90792, 90832, 90834, 90837, 90845, 92002, 92004, 92012, 92014, 96150, 96151, 96152, 97165, 97166, 97167, 97168, 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, 99341, 99342, 99343, 99344, 99345, 99347, 99348, 99349, 99350
OR
At least one visit during the measurement period (CPT/HCPCS):
92521, 92522, 92523, 92524, 92540, 92557, 96160, 96161, 92625, 99385, 99386, 99387, 99395, 99396, 99397, 99401, 99402, 99403, 99404, 99411, 99412, 99420, 99429, G0438, G0439

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason)

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.)*
Exceptions are used to remove a patient from the denominator of a performance measure when the patient does not receive a therapy or service AND that therapy or service would not be appropriate due to patient-specific reasons. The patient would otherwise meet the denominator criteria. Exceptions are not absolute, and are based on clinical judgment, individual patient characteristics, or patient preferences. The PCPI exception methodology uses three categories of reasons for which a patient may be removed from the denominator of an individual measure. These measure exception categories are not uniformly relevant

across all measures; for each measure, there must be a clear rationale to permit an exception for a medical, patient, or system reason. Examples are provided in the measure exception language of instances that may constitute an exception and are intended to serve as a guide to clinicians. For measure 0028, exceptions may include medical reasons for not screening for tobacco use (eg, limited life expectancy, other medical reason). Although this methodology does not require the external reporting of more detailed exception data, the PCPI recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. The PCPI also advocates the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

For Administrative Claims/Registry:
CPT Category II code with modifier 4004F-1P: Documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason)

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 national recommendations put forth by the IOM and NQF to standardize the collection of race and ethnicity data, PCPI encourages 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
4. From the patients who did not meet the numerator criteria, determine if the provider has documented that the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: documentation of medical reason(s) for not screening for tobacco use (eg, limited life expectancy, other medical reason). If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. --Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with valid exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for QI.

If the patient does not meet the numerator and a valid exception is not present, 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 a 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.)*
IF a PRO-PM, 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.
Claims (Only), Claims (Other), Registry

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 is collected.)*
IF a PRO-PM, identify the specific PROM(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)*
Behavioral Health : Outpatient, Clinician Office/Clinic, Home Health, Other
If other: Occupational therapy evaluation, speech and hearing evaluation, ophthalmological services visit

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. The measure is not a composite.

2. Validity – See attached Measure Testing Submission Form
NQF0028_TobaccoTesting_Attachment_Final.doc

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. (Do not remove prior testing information – include date of new information in red.)
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. (Do not remove prior testing information – include date of new information in red.)
Yes

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
No - This measure is not risk-adjusted

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 0028:3225 NQF Project: Behavioral Health

2. RELIABILITY & 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. (evaluation criteria)

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

2a2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

The data are from 301 physicians and other mid-level providers (eg, nurse practitioners, midwives, and Physician Assistants) in a large, urban safety-net network.

The data were collected from a network of community health centers serving primarily low-income and uninsured patients with multiple, complex needs located in the Midwestern US.

The total number of quality events assessed is 13,312. The data are from calendar year 2011.

The site used SQL queries of the EHR to select eligible providers.

REGISTRY – Signal to Noise Ratio Analysis (PQRS)

The data source is Registry data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS). The data are for the time period January 2015 through December 2015 and cover the entire United States.

The total number of physicians reporting on this measure, via the Registry option, in 2015, is 30,033. Of those, 29,949 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 9,391,919 quality events. For this measure, 90.7 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 312.8 for the remaining 9,368,676 events. The range of quality reporting events for 29,949 physicians included is from 7302 to 10. The average number of quality reporting events for the remaining 9.3 percent of physicians that aren't included is 4.05.

There were 9,368,676 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 and remained after exceptions were removed.

CLAIMS – Signal to Noise Ratio Analysis (PQRS)

The data source is Claims data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS). The data are for the time period January 2015 through December 2015 and cover the entire United States.

The total number of physicians reporting on this measure, via the CLAIMS option, in 2015, is 71,445. Of those, 53,326 physicians had all the required data elements and met the minimum number of quality reporting events (10) for a total of 10,177,218 quality events. For this measure, 74.6 percent of physicians are included in the analysis, and the average number of quality reporting events after exceptions are removed is 190.5 for the remaining 10,163,456 events. The range of quality reporting events for 53,326 physicians included is from 3,923 to 10. The average number of quality reporting events for the remaining 25.4 percent of physicians that aren't included is 3.5.

There were 10,163,456 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 and remained after exceptions were removed.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:

Analytic Method

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 eligible provider performance. Reliability at the level of the specific eligible provider is given by:

Reliability = Variance (eligible provider-to-eligible provider) / [Variance (eligible provider-to-eligible provider) + Variance (eligible provider-specific-error]

Reliability is the ratio of the eligible provider-to-eligible provider variance divided by the sum of the eligible provider-to-eligible provider variance plus the error variance specific to a eligible provider. 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 eligible provider performance.

Reliability testing was performed by using a beta-binomial model. The beta-binomial model assumes the eligible provider performance score is a binomial random variable conditional on the eligible provider'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, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per eligible provider who met the threshold for inclusion in the analysis. For this measure, the reliability was estimated at 3 different minimum thresholds for inclusion: 10, 20, and 30 events.

REGISTRY, CLAIMS – Signal to Noise Ratio analysis (PQRS)

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, at the minimum number of quality reporting events for the measure and at the mean number of quality reporting events per physician.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:

Testing results –

The reliability was estimated for eligible providers who met the minimum number of quality reporting events for inclusion in the analysis. We conducted the analysis using a minimum of 10, 20, and 30 events. The number of eligible providers eligible for inclusion went from 175 when the threshould was 10 events to 147 when the threshold was 20 events to 126 when the threshold was 30. The average number of quality reporting events for eligible providers included at 10 events is 76.1 for a total of 13,312 events; at 20 events is 87.8 for a total of 12,908 events and at 30 events is 98.4 for a total of 12,403 events . The range of quality reporting events for eligible providers included is from 389 to 10, 389 to 20, and 389 to 30.

For this measure, the reliability at the minimum level of quality reporting events varied between the three minimum thresholds for inclusion. At 10 quality events the reliability was 0.46. Increasing the threshold for inclusion to 20 events raises the reliability to 0.61 and at 30 events the reliability is 0.69. The reliability at the average number of quality reporting events was stable in the 0.86 to 0.88 range.

1) Minimum number of events

2) Average number of events

3) Number of eligible providers meeting threshold

4) Reliability at minimum number of events

5) Reliability at average number of events

1)	2)	3)	4)	5)
10	76.1	175	0.46	0.86
20	87.8	147	0.61	0.87
30	98.4	126	0.69	0.88

This measure has high and stable reliability when evaluated at the average number of quality events. When increasing the minimum threshold for inclusion, the reliability evaluated at that threshold increases.

Data analyses were conducted by using SAS/STAT software, version 8.2 (SAS Institute, Cary, North Carolina).

REGISTRY – Signal to Noise Ratio analysis (PQRS)

This measure has 0.78 reliability when evaluated at the minimum level of quality reporting events and 0.99 reliability when evaluated at the average number of quality events.

Reliability at the minimum level of quality reporting events is moderate. Reliability at the average number of quality events is very high.

CLAIMS – Signal to Noise Ratio analysis (PQRS)

This measure has 0.71 reliability when evaluated at the minimum level of quality reporting events and 0.97 reliability when evaluated at the average number of quality events.

Reliability at the minimum level of quality reporting events is moderate. Reliability at the average number of quality events is very high.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:

The measure focuses on routine tobacco screening for all adults and tobacco cessation interventions for those who use tobacco products.

Clinical practice guidelines from the U.S. Public Health Service (PHS) and recommendations statements from the U.S. Preventive Services Task Force (USPSTF) recommend that clinicians ask all adults about tobacco use and provide tobacco cessation interventions for those who use tobacco products. The PHS guideline noted that the majority of clinician attention and research in the field has focused on the treatment and assessment of smoking. Nevertheless, they indicated that "[t]he interventions found to be effective in this Guideline have been shown to be effective in a variety of populations. In addition, many of the studies supporting these interventions comprised diverse samples of tobacco users. Therefore, interventions identified as effective in this Guideline are recommended for all individuals who use tobacco, except when medication use is contraindicated or with specific populations in which medication has not been shown to be effective (pregnant women, smokeless tobacco users, light smokers, and adolescents)."

As a basis for their recommendations, the USPSTF reviewed new evidence in the PHS guideline.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

An expert panel was asked to empirically assess face validity of the measure. This panel consists of 30 members, whose specialties include: family medicine, internal medicine, geriatric medicine, gastroenterology, general surgery, colon & rectal surgery, infectious

disease, radiology, cardiology, obstetrics & gynecology, emergency medicine, preventive medicine, occupational medicine, nursing, psychology, occupational therapy, chiropractics, dietetics, optometry.

Our expert panel included 30 members including:

Martin C. Mahoney, MD, PhD (Co-Chair) (family medicine)

Stephen D. Persell, MD, MPH (Co-Chair) (internal medicine)

Gail M. Amundson, MD, FACP (internal medicine/geriatrics)

Joel V. Brill MD, AGAF, FASGE, FACG (gastroenterology)

Steven B. Clauser, PhD

Will Evans, DC, Phd, CHES (chiropractic)

Ellen Giarelli, EdD, RN, CRNP (nurse practitioner)

Amy L. Halverson, MD, FACS (colon & rectal surgery)

Kay Jewell, MD, ABHM (internal medicine/geriatrics)

Daniel Kivlahan, PhD (psychology)

Paul Knechtges, MD (radiology)

George M. Lange, MD, FACP (internal medicine/geriatrics)

Trudy Mallinson, PhD, OTR/L/NZROT (occupational therapy)

Nasseer Masoodi, MD (geriatrics)

Jacqueline W. Miller, MD, FACS (general surgery)

Adrienne Mims, MD, MPH (geriatric medicine)

G. Timothy Petito, OD, FAAO (optometry)

Rita F. Redberg, MD, MSc, FACC (cardiology)

Barbara Resnick, PhD, CRNP (nurse practitioner)

Sam JW Romeo, MD, MBA

Carol Saffold, MD (obstetrics & gynecology)

Robert A. Schmidt, MD (radiology)

Samina Shahabbudin, MD (emergency medicine)

James K. Sheffield, MD (health plan representative)

Arthur D. Snow, MD, CMD (family medicine/geriatrics)

Richard J. Snow, DO, MPH

Brian Svazas, MD, MPH, FACOEM, FACPM (preventive medicine)

David J. Weber, MD, MPH (infectious disease)

Deanna R. Willis, MD, MBA, FAAFP (family medicine)

Charles M. Yarborough, III, MD, MPH (occupational medicine)

The expert panel included 10 members. Panel members were comprised of the newly convened PCPI Preventive Care Technical Expert Panel did not participate in the original work group.

The list of expert panel members are as follows:

Sandra Dunbar, PHD, RN

Peter Briss, MD, MPH

Yngve Falck, MD

Susan Friedman, MD, MPH

Marc Ghany, MD

Ashley Halle, OTD, OTR/L

Selena Hariharan, MD

Lori Karan, MD

Andrew J Saxon, MD

John Wong, MD

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:

All PCPI performance measures are assessed for content validity by a panel of expert work group members during the development process. Additional input on the content validity of draft measures is obtained through a 30-day public comment period and by also soliciting comments from a panel of consumer, purchaser, and patient representatives convened by the PCPI specifically for this purpose. All comments received are reviewed by the expert work group and the measures adjusted as needed. Other external review groups (eg, focus groups) may be convened if there are any remaining concerns related to the content validity of the measures.

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel (workgroup membership described above) was asked to rate their agreement with the following statement:

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.

Scale 1-5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5= Strongly Agree

Face validity of the measure score as an indicator of quality was systematically assessed as follows.

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement:

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.

Scale 1-5, where 1= Strongly Disagree; 2= Disagree; 3= Neither Agree nor Disagree; 4= Agree; 5= Strongly Agree

---

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's ICD-10 conversion approach was used to identify ICD-10 codes for this measure. 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 two RHIA-credentialed professionals on our staff who review all ICD-10 coding. For measures included in PQRS, 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.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:

The results of the expert panel rating of the validity statement were as follows: N = 17; Mean rating = 4.59 and 94.1% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality

Frequency Distribution of Ratings

1 - 0 (Strongly Disagree)

2 - 1

3 - 0 (Neither Agree nor Disagree)

4 - 4

5 - 12 (Strongly Agree)

Frequency Distribution of Ratings

1 – 1 responses (Strongly Disagree)

2 – 0 responses (Disagree)

3 – 3 responses (Neither Agree nor Disagree)

4 – 4 responses (Agree)

5 – 2 responses (Strongly Agree)

The results of the expert panel rating of the validity statement were as follows: N = 10; Mean rating = 3.6 and 60.0% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Given that the majority of expert panel members agreed that the measure can accurately distinguish good and poor quality, the measure is valid, as specified.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

The current structure of this measure doesn't allow for any exception based on patient preferences or any other reason.

REGISTRY, CLAIMS –Exceptions Analysis (PQRS)

The data source is Registry and Claims data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS).

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:

The current structure of this measure doesn't allow for any exception based on patient preferences or any other reason.

REGISTRY, CLAIMS Exceptions Analysis (PQRS)

Exceptions included documentation of medical reason for not screening for tobacco use. Exceptions were analyzed for frequency and variability across providers.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

The current structure of this measure doesn't allow for any exception based on patient preferences or any other reason.

REGISTRY – PQRS Exceptions Analysis (PQRS)

Amongst the 29,949 physicians with the minimum (10) number of quality reporting events, there were a total of 23,243 exceptions reported. The average number of exceptions per physician in this sample is 0.8. The overall exception rate is 0.2%.

CLAIMS – PQRS Exceptions Analysis (PQRS)

Amongst the 53,326 physicians with the minimum (10) number of quality reporting events, there were a total of 13,762 exceptions reported. The average number of exceptions per physician in this sample is 0.3. The overall exception rate is 0.1%.

Exceptions are necessary to account for those situations when it is not medically appropriate for a patient to have tobacco screening. Exceptions are discretionary and the methodology used for measure exception categories are not uniformly relevant across all measures; for this measure, there is a clear rationale to permit an exception for medical reasons. Rather than specifying an exhaustive list of explicit reasons for exception for this measure, the measure developer relies on clinicians to link the exception with a specific medical reason for the decision to screen for tobacco use.

Some have indicated concerns with exception reporting including the potential for physicians to inappropriately exclude patients to enhance their performance statistics. Research has indicated that levels of exception reporting occur infrequently and are generally valid (Doran et al., 2008), (Kmetik et al., 2011). Furthermore, exception reporting has been found to have substantial benefits: "it is precise, it increases acceptance of [pay for performance] programs by physicians, and it ameliorates perverse incentives to refuse care to "difficult" patients." (Doran et al., 2008).

Although this methodology does not require the external reporting of more detailed exception data, the measure developer recommends that physicians document the specific reasons for exception in patients' medical records for purposes of optimal patient management and audit-readiness. We also advocate for the systematic review and analysis of each physician's exceptions data to identify practice patterns and opportunities for quality improvement.

Without exceptions, the performance rate would not accurately reflect the true performance of that physician. This would result in an increase in performance failures and false negatives. The additional value of increased data collection of capturing an exception greatly outweighs the reporting burden.

References:

Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of pay for performance targets by English Physicians. New Engl J Med. 2008; 359: 274-84.

Kmetik KS, Otoole MF, Bossley H et al. Exceptions to Outpatient Quality Measures for Coronary Artery Disease in Electronic Health Records. Ann Intern Med. 2011;154:227-234

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 Data/Sample *(Description of the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

This measure is not risk adjusted.

This measure is not risk adjusted.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

This measure is not risk adjusted.

This measure is not risk adjusted.

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

Not applicable.

Not applicable.

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
Not applicable.

Not applicable.

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

The data are from 301 physicians and other mid-level providers (eg, nurse practitioners, midwives, and Physician Assistants) in a large, urban safety-net network.

The data were collected from a network of community health centers serving primarily low-income and uninsured patients with multiple, complex needs located in the Midwestern US.

The total number of quality events assessed is 13,312. The data are from calendar year 2011.

The site used SQL queries of the EHR to select eligible providers.

REGISTRY,CLAIMS – Signal to Noise Ratio Analysis (PQRS)

The data source is Registry and Claims data from the PQRS program, provided by the Center for Medicare & Medicaid Services (CMS). The data are for the time period January 2015 through December 2015 and cover the entire United States.

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:

Measures of central tendency, variability, and dispersion were calculated.

REGISTRY,CLAIMS – Signal to Noise Ratio Analysis (PQRS)

Measures of central tendency, variability, and dispersion were calculated.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

The performance rate among the three different analyses is listed below:

1) Minimum Number of Events

2) Mean

3) Median

4) Mode

5) Minimum

6) Maximum

7) Standard Deviation

8) Range

9) Interquartile Range

1)	2)	3)	4)	5)	6)	7)	8)	9)
10		0.911	0.953	1.000	0.500	1.000	0.100	0.500	0.132
20		0.915	0.957	1.000	0.518	1.000	0.089	0.482	0.129
30		0.919	0.960	1.000	0.518	1.000	0.083	0.482	0.125

REGISTRY – Signal to Noise Ratio analysis (PQRS)

Based on the sample of 29,949 included physicians, the mean performance rate is 0.84 the median performance rate is 0.93 and the mode is 1. The standard deviation is 0.23. The range of the performance rate is 1, with a minimum rate of 0 and a maximum rate of 1. The interquartile range is 0.17 (0.82 – 0.99).

CLAIMS – Signal to Noise Ratio analysis (PQRS)

Based on the sample of 53,326 included physicians, the mean performance rate is 0.96 the median performance rate is 1.00 and the mode is 1. The standard deviation is 0.11. The range of the performance rate is 1, with a minimum rate of 0 and a maximum rate of 1. The interquartile range is 0.03 (0.97 – 1.00).

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

This test was not performed for this measure.

This test was not performed for this measure.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:

This test was not performed for this measure.

This test was not performed for this measure.

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:

This test was not performed for this measure.

This test was not performed for this measure.

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: We encourage the results of this measure to be stratified by race, ethnicity, gender, and primary language, and have included these variables as recommended data elements to be collected.

Data are not available to complete this testing.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:

The PCPI advocates that performance measure data should, where possible, be stratified by race, ethnicity, and primary language to assess disparities and initiate subsequent quality improvement activities addressing identified disparities, consistent with recent national efforts to standardize the collection of race and ethnicity data. A 2008 NQF report endorsed 45 practices including stratification by the aforementioned variables.(1) A 2009 IOM report "recommends collection of the existing Office of Management and Budget (OMB) race and Hispanic ethnicity categories as well as more fine-grained categories of ethnicity(referred to as granular ethnicity and based on one's ancestry) and language need (a rating of spoken English language proficiency of less than very well and one's preferred language for health-related encounters)."(2)

References:

(1)National Quality Forum Issue Brief (No.10). Closing the Disparities Gap in Healthcare Quality with Performance Measurement and Public Reporting. Washington, DC: NQF, August 2008.

(2)Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. March 2010. AHRQ Publication No. 10-0058-EF. Agency for Healthcare Research and Quality, Rockville, MD. Available at:

http://www.ahrq.gov/research/iomracereport. Accessed May 25, 2010.

Data are not available to complete this testing.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?

(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐

Provide rationale based on specific subcriteria:
If the Committee votes No, STOP

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**.
Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).
Although the claims data is captured electronically with encounter codes for the denominator and CPT II codes for the numerator, registry implementation may vary.

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 a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.
We have not identified any areas of concern or made any modifications as a result of testing and operational use of the measure in
relation to data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, and other feasibility issues unless otherwise noted.

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, eg, 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(R) Foundation (PCPI[R]) or the American Medical Association (AMA). Neither the American Medical Association (AMA), nor the AMA-convened Physician Consortium for Performance Improvement(R) (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 Physician Quality Reporting System (PQRS) http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/MeasuresCodes.html Public Health/Disease Surveillance Million Hearts Initiative http://millionhearts.hhs.gov/

4a.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 Quality Reporting System (PQRS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)
PQRS is a national reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). The program provides an incentive payment to practices with EPs (identified on claims by their individual National Provider Identifier [NPI] and Tax Identification Number [TIN]). EPs satisfactorily report data on quality measures for covered Physician Fee Schedule (PFS) services furnished to Medicare Part B Fee-for-Service (FFS) beneficiaries (including Railroad Retirement Board and Medicare Secondary Payer). Beginning in 2015, the program also applies a payment adjustment to EPs who do not satisfactorily report data on quality measures for covered professional services in 2013.
Source: http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/index.html CMS has implemented
a phased approach to public reporting performance information on the Physician Compare Web site.

CMS has implemented a phased approach to publicly reporting performance information on the Physician Compare Web Site. Beginning with PQRS 2014 reporting, this measure is one of 14 group practice PQRS measures reported via the Web interface that are currently available for public reporting. This measure is also one of 6 individual EP PQRS measures reported via claims that are currently available for public reporting. CMS also announced through rulemaking their plans to make all PQRS individual EP level PQRS measures available for public reporting annually, including making the 2016 PQRS individual EP level data available for public reporting on Physician Compare in 2017. Beginning in 2017, the Merit-based Incentive Payment System (MIPS) consolidates PQRS and other existing quality reporting programs. This measure has been finalized as an individual quality measure available for MIPS reporting in 2017.

Million Hearts
Million Hearts™ is a national initiative to prevent 1 million heart attacks and strokes in the U.S. over the next 5 years. Launched by the Department of Health and Human Services (HHS) in September 2011, it aligns existing efforts, as well as creates new programs, to improve health across communities and help Americans live longer, more productive lives. The Centers for Disease Control and Prevention (CDC) and Centers for Medicare & Medicaid Services (CMS), co-leaders of Million Hearts™ within HHS, are working alongside other federal agencies and private-sector organizations to make a long-lasting impact against cardiovascular disease.

The Million Hearts® Clinical Quality Measures (CQM) Dashboard is designed to display quality reporting measures focused on the Million Hearts®ABCS (Aspirin when appropriate, Blood pressure control, Cholesterol management, and Smoking cessation) and is based on information from the following available data systems, where possible.

HRSA UDS - Health Resources and Services Administration Uniform Data System
NCQA HEDIS - National Committee for Quality Assurance Healthcare Effectiveness Data and Information Set
CMS PQRS - Centers for Medicare & Medicaid Services Physician Quality Reporting System

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Not applicable

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
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.

4b. **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.
Although the PQRS program has demonstrated increasing performance rates over time which would indicate progress on improvement, it's important to note that the percentage of eligible professional reporting on PQRS measures overall and on this measure, in particular, continues to grow but remains low. In 2014, for example, only 21.7% of eligible professionals reported on the measure. As a result, performance rates may not be nationally representative.

Additionally, 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.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **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 measure.

4c.2. **Please explain any unexpected benefits from implementation of this measure.**

We are not yet aware of any unexpected benefits related to this measure.

4d1.1. **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, 32 individuals from a diverse group of specialties including family medicine, internal medicine, geriatric medicine, gastroenterology, general surgery, nursing, and psychology participated on the measure development work group.

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 works 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.

4d1.2. **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 4d1.1 above.

4d2.1. **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.

4d2.2. **Summarize the feedback obtained from those being measured.**
The majority of comments received during the public comment period were supportive and approving of the broad nature of the measure, its potential for public health impact and patient outcomes. There were some specific comments requesting consideration of a lower age range for the measure and adding a medical reason exception for patients with limited life expectancy.

The majority of feedback from implementers seeks to have the PCPI clarify what qualifies and does not qualify as meeting the measure. More recently, many implementers wanted to understand how the measure addresses electronic nicotine delivery systems (ENDS).

4d2.3. **Summarize the feedback obtained from other users**
See summary in 4d2.2 above.

4d.3. **Describe how the feedback described in 4d.2 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.**
At the time of original development, the expert work group decided not to adjust the age range as it was developed to align with the USPSTF's recommendation for adults. The latter comment regarding the medical reason exception was incorporated into the final version of the measure.

As a result of implementation feedback, a brief definition of cessation intervention has been added to the measure. Guidance has been provided to explain the omission of ENDS from the measure and the rationale for doing so.

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)**
0027 : Medical Assistance With Smoking and Tobacco Use Cessation
1651 : TOB-1 Tobacco Use Screening
1654 : TOB - 2 Tobacco Use Treatment Provided or Offered and the subset measure TOB-2a Tobacco Use Treatment
1656 : TOB-3 Tobacco Use Treatment Provided or Offered at Discharge and the subset measure TOB-3a Tobacco Use Treatment at Discharge

2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
2803 : Tobacco Use and Help with Quitting Among Adolescents

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?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Related measures have differing target populations and/or levels of measurement from the PCPI's Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention measure. 0028 focuses on routine tobacco screening for all adults and tobacco cessation interventions for those who use tobacco products and is intended to assess clinician level performance towards these objectives. The cessation intervention required by the PCPI measure includes brief counseling and/or pharmacotherapy in light of the strong support for these interventions in the guidelines and the feasibility of implementing these practices as part of routine care. Measure 0027 is a patient survey measure assessing health plan performance and includes one additional component of the cessation intervention beyond our measure (ie, discussion of methods or strategies other than medication). Measures 1651, 1654 and 1656 assess hospital level performance at providing tobacco use and treatment to patients being discharged from hospitals. Measure 2803 is focused on assessing clinical level performance on tobacco cessation counseling among adolescents. Finally, measure 2600 represents an adaptation of the PCPI measure and is limited to a subset of the population of patients with serious mental illness.

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.)
No competing measures.

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 Foundation
Co.2 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-464-5524-
Co.3 **Measure Developer if different from Measure Steward:** PCPI Foundation

Co.4 **Point of Contact:** Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-464-5524-

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.

Gail M. Amundson, MD, FACP (internal medicine/geriatrics)
Joel V. Brill MD, AGAF, FASGE, FACG (gastroenterology)
Steven B. Clauser, PhD
Will Evans, DC, Phd, CHES (chiropractic)
Ellen Giarelli, EdD, RN, CRNP (nurse practitioner)
Amy L. Halverson, MD, FACS (colon & rectal surgery)
Alex Hathaway, MD, MPH, FACPM
Charles M. Helms, MD, Phd (infectious disease)
Kay Jewell, MD, ABHM (internal medicine/geriatrics)
Daniel Kivlahan, PhD (psychology)
Paul Knechtges, MD (radiology)
George M. Lange, MD, FACP (internal medicine/geriatrics)
Trudy Mallinson, PhD, OTR/L/NZROT (occupational therapy)
Elizabeth McFarland, MD (radiology)
Jacqueline W. Miller, MD, FACS (general surgery)
Adrienne Mims, MD, MPH (geriatric medicine)
Sylvia Moore PhD, RD, FADA (dietetics)
G. Timothy Petito, OD, FAAO (optometry)
Rita F. Redberg, MD, MSc, FACC (cardiology)
Barbara Resnick, PhD, CRNP (nurse practitioner)
Sam JW Romeo, MD, MBA (family practice)
Carol Saffold, MD (obstetrics & gynecology)
Robert A. Schmidt, MD (radiology)
Samina Shahabbudin, MD (emergency medicine)
James K. Sheffield, MD (health plan representative)
Arthur D. Snow, MD, CMD (family medicine/geriatrics)
Richard J. Snow, DO, MPH
Brooke Steele, MD
Brian Svazas, MD, MPH, FACOEM, FACPM (preventive medicine)
David J. Weber, MD, MPH (infectious disease)
Deanna R. Willis, MD, MBA, FAAFP (family medicine)
Charles M. Yarborough, III, MD, MPH (occupational medicine)

PCPI measures are developed through cross-specialty, multi-disciplinary work groups. All medical specialties and other health care professional disciplines participating in patient care for the clinical condition or topic under study must be 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. 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. All work groups have at least two co-chairs who have relevant clinical and/or measure development expertise and who are responsible for ensuring that consensus is achieved and that all perspectives are voiced.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2001

Ad.3 **Month and Year of most recent revision:** 11, 2015

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?** 11, 2016

The Measures are not clinical guidelines, do not establish a standard of medical care, and have not been tested for all potential applications.

The Measures, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, eg, 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.

AMA and PCPI encourage use of the Measures by other health care professionals, where appropriate.

THE MEASURES AND SPECIFICATIONS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND.

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. The AMA, the PCPI and its members and former members of the AMA-PCPI disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

Ad.7 **Disclaimers:** See copyright statement above.

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 #: 3229
Measure Title: Patient Panel Adult Smoking Prevalence
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: Percentage of adults (age 18 years or older) who are tobacco smokers at time of most recent encounter during the measurement period.
Developer Rationale: Health behaviors such as smoking are critical contributors to multiple poor health outcomes, and there is a large body of evidence that health behaviors are amenable to modification by practitioners in the clinical care sector. Despite declines in use since its peak in 1965, tobacco consumption, and cigarette smoking in particular, remains the single most preventable cause of disease and death in the US (HHS 2014). As of 2015, an estimated 36.5 million (15.1%) adults endorsing cigarette smoking. (CDC 2016).

To date, quality performance measures associated with smoking have focused on clinical processes related to risk assessment and appropriate follow-up activities for patients as needed. A transition to outcome measures would move the focus beyond just screening and referrals, but also on identifying and executing effective strategies for risk reduction, including both clinical interventions and clinical-community partnerships. When combined with health system-specific changes that may be needed to support efficient workflow and data feedback loops to facilitate continuous learning, current process measures for tobacco cessation may fall short of recognizing all the variation and nuance potentially needed for long-term success. Outcome measures offer an alternative that would allow for evaluation of context-specific interventions, fit well into an environment of constant adaptation and continuously emerging properties, and have applicability in other population-based initiatives targeting optimized levels of health system efficiency and cultural sensitivity. Outcome measures offer a mechanism to examine tailored approaches to operationalizing the 5 As (ask, advise, assess, assist, arrange), and to assess the impact of collaboration with external organizations or agencies representing the public health or community sectors. In shifting to outcomes, the measure focus would emphasize the end result rather than a checklist of procedural steps. Thus, the use of outcome measures could encourage the development and refinement of effective innovations best-suited and adapted to a particular context (community, health care system, patient), resulting in population health improvements via patient-centered care.

Citations:

HHS (United States Department of Health and Human Services). (2014) The health consequences of smoking—50 years of progress: A report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf. Accessed 9 September 2015.

CDC (Centers for Disease Control and Prevention). (2016) Current Cigarette Smoking Among U.S. Adults Aged 18 Years and Older. https://www.cdc.gov/tobacco/campaign/tips/resources/data/cigarette-smoking-in-united-states.html. Accessed 14 February 2016.

CMS (Centers for Medicare and Medicaid Services) Medicaid (2015) Tobacco Cessation. Accessed 9 September 2015 from: http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Benefits/Tobacco.html.

Numerator Statement: Adult patients identified as smokers.
Denominator Statement: Adult patients who had a qualifying encounter with a provider during the measurement period AND were identified as either as smokers or non-smokers within 24 months of the end of the measurement period.
Denominator Exclusions: Adult patients were excluded if their smoking status (either as a smoker or a non-smoker) was missing.

Measure Type: Intermediate outcome
Data Source: Electronic Clinical Data: Electronic Clinical Data
Level of Analysis: Clinician : Individual

New Measure - Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *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.
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:
This intermediate clinical outcome calculates the percentage of patients identified as "smokers".

- The developer provided a clinical practice guideline with three recommendations from the Treating Tobacco Use and Dependence 2008 guideline **(Strength of Evidence: Strong):**
  - All patients should be asked if they use tobacco and should have their tobacco use status documented on a regular basis. Evidence has shown that clinic screening systems, such as expanding the vital signs to include tobacco use status or the use of other reminder systems such as chart stickers or computer prompts, significantly increase rates of clinician intervention (p. 77).
  - All *physicians* should strongly advise every patient who smokes to quit because evidence shows that physician advice to quit smoking increases abstinence rates (p. 82).
  - Treatment delivered by a variety of clinician types increases abstinence rates. Therefore, all clinicians should provide smoking cessation interventions (p. 87).
- The developer also provided the following USPSTF recommendations from the Behavioral Counseling and Pharmacotherapy Interventions for Tobacco Cessation in Adults, Including Pregnant Women **(Grading of Evidence A)**:

- The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and US Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco.
- The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco.

- The developer provided a systematic review (SR) of the body of evidence concluding that clinicians can substantially increase the odds that a patient will try to quit smoking and will be successful in quitting if clinicians screen all patients for tobacco use and offer evidence-based cessation treatments to those who use tobacco.
- The developer provided a summary of the Quantity, Quality, and Consistency (QQC) of the systematic review of the body of evidence.

Exception to evidence: N/A

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the intermediate clinical outcome that is being measured?*

Guidance from the Evidence Algorithm: Intermediate clinical outcome measure with SR and grading of the body of evidence (Box 3) → QQC summary from the SR of the body of evidence (Box 4) → The SR concludes QQC is High (high certainty that the net benefit is substantial) → High

Preliminary rating for evidence: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement **and 1b.** Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- The developer calculated the following provider-level smoking prevalence rates using PQRS EHR QRDA I data from 2014:

Year	**2014**
# of providers	382
# of patients	72,478
Mean	13.2
Median	9.6
Std Dev	12.0
IQR	12.1
Min, Max	0.0, 69.2
10th Percentile	2.6
20th Percentile	4.6

30th Percentile	5.9
40th Percentile	7.7
60th Percentile	11.9
70th Percentile	14.9
80th Percentile	19.6
90th Percentile	28.2

- Lower values are better as they reflect a lower prevalence of smoking at the provider level.
- The developer also provided the following smoking prevalence data from the literature:
  - As of 2015, an estimated 36.5 million (15.1%) adults were still smoking cigarettes (CDC 2016).
  - Receiving brief advice to quit is associated with a 30% increase in the number of smokers who quit, and cessation programs using either therapy, medication, or a combination of both have success rates between 10%-30% (Gorin and Heck 2004; Katz 2004; Leif Associates 2012; NCQA 2014). Because even brief interventions have been shown to be effective, providers' continued assessment and provision of education can facilitate important opportunities for patients to quit (Fiore 2000; AAMC 2007).
  - The developer noted that performance on the existing PQRS tobacco screening and intervention process measure has demonstrated providers are engaging with patients on this topic. According to the 2014 PQRS Experience Report, average performance on this measure was 81.6% in 2011 and has increased over time to 88.9% in 2014.

Disparities:

- The developer calculated the following smoking prevalence rates stratified by age, race, ethnicity and sex using PQRS EHR QRDA I data from 2014:

	# of patients	% of Total (72,478)	Smoking Prevalence Rate
Age			
Less than 65 years	25,461	35.1%	22.7%
65 to 69 years	12,822	17.7%	11.2%
70 to 74 years	11,470	15.8%	8.3%
75 to 79 years	9,549	13.2%	5.8%
80 to 84 years	7,093	9.8%	4.6%
85 years or older	6,083	8.4%	3.2%
Sex			
Male	29,425	40.6%	14.9%
Female	43,045	59.4%	11.3%
Unknown	8	0.01%	25.0%
Race			
American Indian or Alaskan Native	63	0.1%	25.4%
Asian	2,063	2.9%	5.4%
Black	3,099	4.3%	18.5%
Native Hawaiian or Pacific Islander	24	0.03%	20.8%
White	41,727	57.6%	11.9%
Other	23,177	32.0%	14.2%
Unknown	2,325	3.2%	11.8%
Ethnicity			

Hispanic or Latino	1,849	2.6%	8.9%
Not Hispanic or Latino	67,868	93.6%	12.9%
Unknown	2,761	3.8%	11.1%

- The developer provided additional data on disparities from the literature.

Questions for the Committee:

- *Does the data demonstrate overall less-than-optimal performance and variation in smoking prevalence rates across providers despite an increase on the performance of the existing PQRS tobacco screening and intervention process measure?*
- *Is there opportunity for improvement in decreasing smoking rates?*
- *Do you agree the data demonstrates a disparity in care for various populations?*

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 to Support Measure Focus

Comments:

**It would seem that there are other stronger more robust measures available that address this issue. Yes, identification of smoking status and prevalence are important, but compared to 0027, seem relatively rudimentary. In any case, clear linkage in a causal pathway.
**The evidence support a move to a population based outcome measure for tobacco cessation is good.
**Developers provide a detailed diagram of the linkage between the five A's (Ask, Advise, Assess, Assist, and Arrange) and abstinence from smoking. Numerous studies provide evidence that physician advice positively impacts abstinence from smoking behavior. The developer provided a systematic review of evidence from the USPSTF, and a clinical practice guideline recommendation. The USPSTF evidence included 54 systematic reviews and the clinical guideline body of evidence included randomized control trials and 11 meta-analyses.

1b. Performance Gap

Comments:

**Yes, despite the ongoing efforts.
**The developers provide evidence of significant variation in smoking prevalence across providers, suggesting a significant performance gap in assessing smoking behavior of patients. 1 in 4 adults in the U.S smokes tobacco of one sort or another. The prevalence of tobacco smoking is higher in certain racial groups, persons younger than 65, lower SES, and Medicaid recipients.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): electronic clinical data – this is not an eMeasure.

Specifications:

- The level of analysis is at the individual clinician-level.
- The setting of care includes the outpatient and clinician office/clinic setting, but the developer also notes that PQRS providers may include additional settings: Speech and Hearing Evaluation, Occupational Therapy Evaluation, and Ophthalmological Visits.
- The measurement period is one year.

- A lower score indicates better quality.
- The numerator includes "adult patients identified as smokers as of the last qualifying encounter during the measurement period."
- The denominator includes "adult patients who had a qualifying encounter with a provider during the measurement period AND were identified as either as smokers or non-smokers within 24 months of the end of the measurement period.
- NOTE: The age range for adults and 'qualifying encounter' are not defined here for either the numerator or denominator. However, developer provides a supplemental data dictionary.
- The measure excludes all patients who do not have a smoking status documented, for any reason. NOTE: In the exclusions section of the testing attachment, the developers state that patients with limited life expectancy or a medical reason not to be screened also are excluded from the measure.
- A calculation algorithm is provided.
- The measure is not risk adjusted.

Questions for the Committee:

- *Are all the data elements (e.g., numerator, denominator, exclusions) clearly defined?*
- *Are all appropriate codes included?*
- *Is the calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing, Testing attachment
Maintenance measures – less emphasis if no new testing data provided

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.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- Data used for testing included CY2014 PQRS data reported via the EHR QRDA-1 (i.e., individual patient) reporting option.
  - These data include information from 382 eligible professionals (EPs) and 72,478 patients.
  - Eligible professionals included in this data sample are those who reported smoking status for at least 10 patients and for at least 50% of all patients.
  - It is not clear if the measure specifications limit calculation of the measure to EPs with at least 10 patients and who report smoking status on at least 50% of all patients. If not, then testing has not been conducted for the measure *precisely* as specified. Often, reliability is lower when sample size (i.e., number of patients) is lower. Therefore, limiting the testing sample to EPs with at least 10 patients likely would result in higher reliability estimates than would be seen otherwise. However, it is not clear how limiting the testing sample to EPs who report smoking status on at least 50% of all patients would affect reliability estimates.
- Empirical testing of the performance measure score was conducted via a signal-to-noise analysis using the beta-binomial model. This is an appropriate method for testing reliability.
  - A signal-to-noise analysis quantifies the amount of variation in performance that is due to differences between providers (as opposed to differences due to measurement error). Results will vary based on the amount of variation between the providers and the number of patients treated by each provider.
  - This method results in a reliability statistic that ranges from 0 to 1 for each provider. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance.

- A value of 0.7 often is regarded as a minimum acceptable reliability value.

Results of reliability testing:

- Among the 382 providers included in the testing sample:
  - Average reliability=0.899
  - Median reliability=0.946
  - 25th percentile=0.885
  - 75th percentile=0.976
- As expected, reliability increased as the number of patients per provider increased.

Smoking Prevalence Reliability for Providers Reporting in 2014 PQRS EHR QRDA-1

Group	Eligible	Patients	Average Reliability
Overall	382	72,478	0.899, 0.946 [0.885, 0.976]
Minimum Provider Size			
≥ 20 patients	364	72,197	0.912, 0.946 [0.890, 0.975]
≥ 30 patients	344	71,722	0.922, 0.946 [0.897, 0.975]
≥ 40 patients	328	71,186	0.926, 0.947 [0.904, 0.975]
≥ 50 patients	314	70,568	0.931, 0.948 [0.910, 0.976]
≥ 100 patients	242	65,341	0.948, 0.957 [0.930, 0.979]
Census Region			
Northeast	54	8,563	0.819, 0.872 [0.758, 0.932]
Midwest	56	9,086	0.903, 0.941 [0.877, 0.974]
South	189	36,196	0.889, 0.927 [0.870, 0.966]
West	83	18,633	0.885, 0.938 [0.863, 0.972]
RUCA			
Urban	351	64,986	0.898, 0.947 [0.885, 0.978]
Rural	31	7,492	0.878, 0.910 [0.825, 0.940]

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Does the developer's decision to limit the testing sample to EPs who report smoking status on at least 50% of all patients make sense? How would this affect reliability estimates?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm
Specifications not precise → Low

The highest possible rating is HIGH.

Preliminary rating for reliability: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

RATIONALE: The definition of "adults" is not provided, the measurement period is not clear (i.e., is it one year or 24-months), the exclusions are not well-defined, and it is not clear whether a smoking status is needed for the denominator and for the numerator. The definition of a qualifying encounter is provided in an external file; while allowed, because there are relatively few codes, these would be best included in the actual submission. If the developer clarifies the specifications to the Committee's satisfaction and agrees to modify the submission accordingly, then the measure would be eligible for a HIGH rating, if the Committee is satisfied that the results from the testing sample used for the signal-to-noise reliability analysis adequately reflects the measure as specified.

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Specification not completely consistent with evidence N/A

Question for the Committee:

o *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity**

☒ **Empirical validity testing of the measure score**

Validity testing method:

- Face validity of the performance measure score was assessed by the developer's Health Behaviors Technical Expert Panel, a group of 9 individuals with a variety of expertise and stakeholder perspectives.
  - Eight of the 9 TEP members indicated their agreement with the following statement: *"The scores obtained from the measure as specified will provide an accurate reflection of population health and can be used to distinguish good and poor population health"*.
  - NQF guidance indicates that the assessment of face validity of the measure score as an indication of quality is an acceptable method for measure validation if systematically assessed by recognized experts and ***explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.*** This face validity assessment does not meet NQF requirements**..**
- Developers conducted construct validation testing of the measure scores via three separate analyses. These were appropriate methods of score-level validation.
  - The developers hypothesized that EPs with higher rates of screening and intervention for tobacco use in 2014 would have lower smoking prevalence among their patients in 2015.
  - Data used for testing included CY2014 and CY2015 PQRS data reported via the EHR QRDA-1 option. These data included information from 106 EPs who reported smoking status for at least 10 patients in 2014 and 2015 and who had a 50% or greater reporting rate on the screening and intervention measure in 2014. This testing sample included 13,914 patients in 2014 and 17,742 patients in 2015.
  - For their first analysis, developers correlated CY2014 results from a tobacco use screening and intervention measure (#0028) to results from this measure (smoking prevalence) for CY2015.
  - For their second analysis, developers grouped the 106 EPs according to tertiles of results of the screening and intervention measure for CY2014. They then calculated the 2015 smoking prevalence results for each of the three groups of EPs.
  - For their third analysis, using a simple regression model, developers estimated the association between clinician-level screening and intervention in 2014 with 2015 smoking prevalence, while controlling for 2014 smoking prevalence.

Validity testing results, construct validation:

- Correlation analysis: correlation coefficient = -0.29, p=0.002. These results, which show a statistically significant negative correlation between tobacco screening and intention in one year and smoking prevalence in the next year, support the developer's hypothesis.
- Tertile analysis: EPs with higher rates of tobacco use screening and intervention in 2014 had lower rates of smoking prevalence in 2015. These results support the developer's hypothesis.

2015 Smoking Prevalence by 2014 Tobacco Use Screening and Intervention

2014 Tobacco Use Screening and Intervention	Eligible Professionals	Mean 2015 Smoking Prevalence (%)
38.2% to 58.7% of EP's patients	36	15.3
58.7% to 75.9% of EP's patients	35	12.9
75.9% to 96.0% of EP's patients	35	10.4

- Regression analysis: These results show a statistically significant negative association between 2014 tobacco use screening/intervention rates and 2015 smoking prevalence, after controlling for the 2014 smoking prevalence rate. Specifically, for each percentage point increase in the 2014 screening/intervention rate, there was a 0.2% decrease in smoking prevalence in 2015 (or, as stated by the developer, for each 10 percentage point increase in tobacco use screening/intervention in 2014, there was a 2% decrease in smoking prevalence in 2015). *[NOTE that the developer seemingly did not provide estimated results for the 2014 smoking prevalence variable]*

Regression Model Results

Model Parameter	Estimate [Standard Error]	P-value
Intercept	26.4 [4.4]	<0.0001
Slope	-0.199 [0.064]	0.002

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- According to the specifications, this measure excludes all patients who do not have a smoking status documented, for any reason. It is not clear whether there are additional exclusions for the measure.
- Using CY2014 PQRS data reported via the EHR QRDA-1 reporting option, the developers indicate the number and percentage of patients from the full testing sample that would be excluded from the measure. Specifically, from among the 104,395 patients with encounters with the 382 EPs in 2014:
  - 4,207 patients (4.0%) were excluded due to "specified reasons" (i.e., limited life expectancy or medical reason, as well as age<18).
  - 27,710 patients (26.5%) were excluded due to missing smoking status.
- The developer notes that the high percentage of "missingness" due to unknown smoking status could alter the measure results (i.e., the rate would be 9.2% if excluded patients were non-smokers and would be 36.9% if excluded patients were smokers).

Questions for the Committee:

- *Is it clear which patients would be excluded from the measure?*
- *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)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ Yes ☒ No

- The developer states that this measure "*is unintended to support population health improvement efforts at actionable levels, such as providers. In accordance with the goal of population health improvement, the proposed smoking prevalence measure will not be risk-adjusted*".
- For outcome measures, NQF expects both a rationale and analysis to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities (in this case, for individual clinicians). However, the developer did not provide the required analyses.
- Because the developer has chosen not to risk-adjust this measure, they did not respond to the questions regarding inclusion of socio-demographic factors in the risk-adjustment approach.

Questions for the Committee:

- *Is there any evidence that suggests a need to risk-adjust this measure?*
- *Do you agree with the developer's decision not to risk-adjust the measure?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- To assess the ability to identify meaningful differences in results between EPs, the developers calculated the number and percentage of EPs who had results that were statistically significantly above or below the "national average" (or, more precisely, above or below the average across the EPs in the CY2014 PQRS dataset), both overall and within size strata.
- The data used for this analysis included CY2014 PQRS data reported via the EHR QRDA-1 reporting option (382 EPs and 72,478 patients).
- The distributional statistics provided indicate quite a bit of variation in measure results across EPs.

Distribution of Smoking Prevalence, by EP (PQRS: EHR QRDA-1)

Year	*EPs*	*Mean*	*Median*	*Mode*	*Std Dev*	*25th Pctl*	*75th Pctl*	*IQR*	*Min*	*Max*	*Range*
2014	382	13.2	9.6	0.0	12.0	5.3	17.4	12.1	0.0	69.2	69.2

- Approximately 15.2% of EPs had results that were statistically significantly higher than the EP average and 13.1% had results that were statistically significantly lower than the EP average.

Statistically significant differences compared to national average

Categories	EPs	Percent of EPs	Smoking Prevalence (mean)
Better than expected	**58**	**15.2**	**2.8**
<50 patients	11	16.2	0
50-99 patients	6	8.3	1.3
100-249 patients	18	11.3	2.7
≥250 patients	23	27.7	4.5
As expected	**274**	**71.7**	**11.2**
<50 patients	50	73.5	15.2
50-99 patients	57	79.2	10.5
100-249 patients	121	76.1	9.9
≥250 patients	47	56.6	11.5
Worse than expected	**50**	**13.1**	**36.0**
<50 patients	7	10.3	43.8
50-99 patients	9	12.5	35.0

100-249 patients	20	12.6	37.4
≥250 patients	13	15.7	31.5

- The developers also provided analysis to inform the setting of minimal sample sizes for the measure if comparing to an overall mean. NOTE, however, that the measure does not specify use of a particular minimum sample size.

Question for the Committee:

o *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable, as only one data source and calculation methodology is specified.

2b7. Missing Data

- The developers presented measure results from the CY2014 PQRS QCDR-1 data with and without restricting the measure to those EPs who report smoking status for at least half of their patients. They interpret the results (non-similarity to survey results from Medicare Advantage beneficiaries) as suggesting that the measure may be biased if the EPs are not restricted.
  - o However, the developers do not explain why they believe that EPs who report on the screening/intervention measure via the QCDR-1 option (and their patients) are similar to EPs and their Medicare Advantage patients.
  - o Moreover, as already noted, this measure is not specified to include a minimum sample size or to restrict EPs who are included in the measure.
  - o This analysis does illustrate that specifying such a restriction likely would substantially reduce the number of EPs eligible for the measure.

Guidance from the Validity Algorithm

Specifications consistent with evidence (Box 1) → Threats to validity not completely assessed: need analysis to demonstrate risk adjustment is not needed (Box 2) → Insufficient

The highest possible rating is HIGH.

Preliminary rating for validity: ☐ High ☐ Moderate ☐ Low ☒ Insufficient

RATIONALE: Analysis needed to demonstrate that risk adjustment is not needed for this measure.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

Comments:

**Pretty straight forward and clear

**The Denominator only includes patients who were screened for tobacco use.How do you control for providers who are not practicing universal screening and therefore not capturing the tobacco use status of their patient population.

Currently the denominator does not include patients seen in a SUD setting. Individuals with SUD have high rates of tobacco use and cessation efforts can be very successful in the context of SUD treatment, since tobacco use is an addiction.

Is there a need to risk adjust for groups of individuals who may be in the early stages of trying to quit--considering the evidence that, on average, quitting can take 8 attempts.

**The numerator and denominator are clearly specified. Codes have been created for smoking/non-smoking. There is no risk adjustment made for this measure. Given the simplicity of this measure, it should be implemented consistently across various sites with few problems.

2a2. Reliability Testing

Comments:
**Yes, reliable.
**The reliability testing was conducted on 72,478 patients from 382 EP in all regions of the United States. This appears to be of adequate scope to generalize to most settings across the United States. Reliability of the measure as reported was high (.89).

2b1. Validity Specifications
Comments:
**Valid.
**The validity specifications are consistent with the evidence the developer presented in 1A.

2b2. Validity Testing
Comments:
**Validity testing was conducted using face and empirical validity testing. In terms of the face validity, there was consensus that the measure would provide an accurate reflection of population health and that it could differentiate between good and poor population health. In terms of the empirical testing, the linear regression results suggested that screening and intervention was positively associated with reductions in the smoking prevalence rates. This suggests that the measure is valid and also can be used for quality improvement purposes in this area.

2b3. Exclusions Analysis
2b4. Risk Adjustment/Stratification for Outcome or Resource Use Measures
2b5. Identification of Statistically Significant & Meaningful Differences In Performance
2b6. Comparability of Performance Scores When More Than One Set of Specifications
2b7. Missing Data Analysis and Minimizing Bias
Comments:
**Reasonable response to the issues.
**There were 27,710 missing cases (26.5%) in the analysis, which could impact the prevalence rate findings in the validity testing. Other exclusions appear to be appropriate. The testing used to show meaningful differences exist in performance suggests that there were quantifiable differences across EPs.

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 generated during the routine delivery of care and available in defined fields in a combination of electronic clinical data.
- The developer reported that no new cost, burden or challenges for providers who use CPT, SNOMEDCT, and HCPCS.
- The developer reported mentioned that CMS may need to establish minimum data reporting standards to fully implement this measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

3a. Byproduct of Care Processes
3b. Electronic Sources
3c. Data Collection Strategy

Comments:
**Yes, clearly feasible.
**The measure does not pose any additional cost or burden to providers. Most electronic sources collect the measure components. This information is collected and used by healthcare professionals in the routine delivery of care. CMS may establish minimum data reporting standards for this measure.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No** ☐ **UNCLEAR**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The developer intends to use the measure for public reporting within the next six years and is considering using this measure in the Medicare Shared Savings Program and CMS Innovation Center models. The developer also intends to submit the measure for review under the Measures Under Consideration (MUC) List for review by NQF's Measure Application Partnership (MAP).

Improvement results N/A

Unexpected findings (positive or negative) during implementation New measure. No unexpected findings reported.

Potential harms None reported.

Vetting of the measure: New measure. None reported.

Feedback: New measure. 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?*
- o *How has the measure been vetted in real-world settings by those being measure or others?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

4a. Accountability and Transparency
4b. Improvement
4c. Unintended Consequences
Comments:
**Useful, perhaps, but really a first step toward improved patient oriented outcomes.
**CMS intends to use the proposed measures for public reporting in 6 years. CMS may consider using the measure in different payment models. The benefits of collecting this information outweigh any unintended consequences

Criterion 5: Related and Competing Measures

Related or competing measures

- 0028/3225/3185 : Preventive Care and Screening: Tobacco Use: Screening and Cessation Intervention
- 1651 (TOB-1): Tobacco Use Screening
- 2600 : Tobacco Use Screening and Follow-up for People with Serious Mental Illness or Alcohol or Other Drug Dependence
- 2803 : Tobacco Use and Help with Quitting Among Adolescents

Harmonization

- The developer notes they have developed this measure based on the specifications of measure 3225. Overall, these measures seem to be harmonized to the degree possible.
- Measure 1651 is a hospital-level measure aimed at screening adults for tobacco use at the facility level of analysis. The measures seem to be mostly harmonized (1651 has exclusions for cognitive impairment and patients on 'comfort measures only.')
- Measure 2600 focuses on specific populations (SMI, AOD) at the health plan level
- Measure 2803 looks at screening and cessation interventions in adolescents at the clinician group/practice level.

Endorsement + Designation

The "Endorsement +" designation identifies measures that exceed NQF's endorsement criteria in several key areas. After a Committee recommends a measure for endorsement, it will then consider whether the measure also meets the "Endorsement +" criteria.

This measure is a candidate for the "Endorsement +" designation IF the Committee determines that it: meets evidence for measure focus without an exception; is reliable, as demonstrated by score-level testing; is valid, as demonstrated by score-level testing (not via face validity only); and has been vetted by those being measured or other users.

Eligible for Endorsement + designation: ☐ **Yes** ☒ **No**

RATIONALE IF NOT ELIGIBLE: New measure – there has been no vetting.

Pre-meeting public and member comments

- None received.

Measure Title: Patient Panel Adult Smoking Prevalence

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: 1/27/2017

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

Subcriterion 1a. Evidence to Support the Measure Focus

The measure focus is a health outcome or is evidence-based, demonstrated as follows:

- Health outcome:[3] a rationale supports the relationship of the health outcome to processes or structures of care.
- Intermediate clinical outcome, Process,[4] or Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence[5] that the measure focus leads to a desired health outcome.
- Patient experience with care: evidence that the measured aspects of care are those valued by patients and for which the patient is the best and/or only source of information OR that patient experience with care is correlated with desired outcomes.
- Efficiency:[6] evidence for the quality component as noted above.

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. 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.
5. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

6. Measures of efficiency combine the concepts of resource use and quality (NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of:

Outcome

☐ Health outcome:

Health outcome includes patient-reported outcomes (PRO, i.e., HRQoL/functional status, symptom/burden, experience with care, health-related behaviors)

☒ Intermediate clinical outcome: Adult Smoking Prevalence

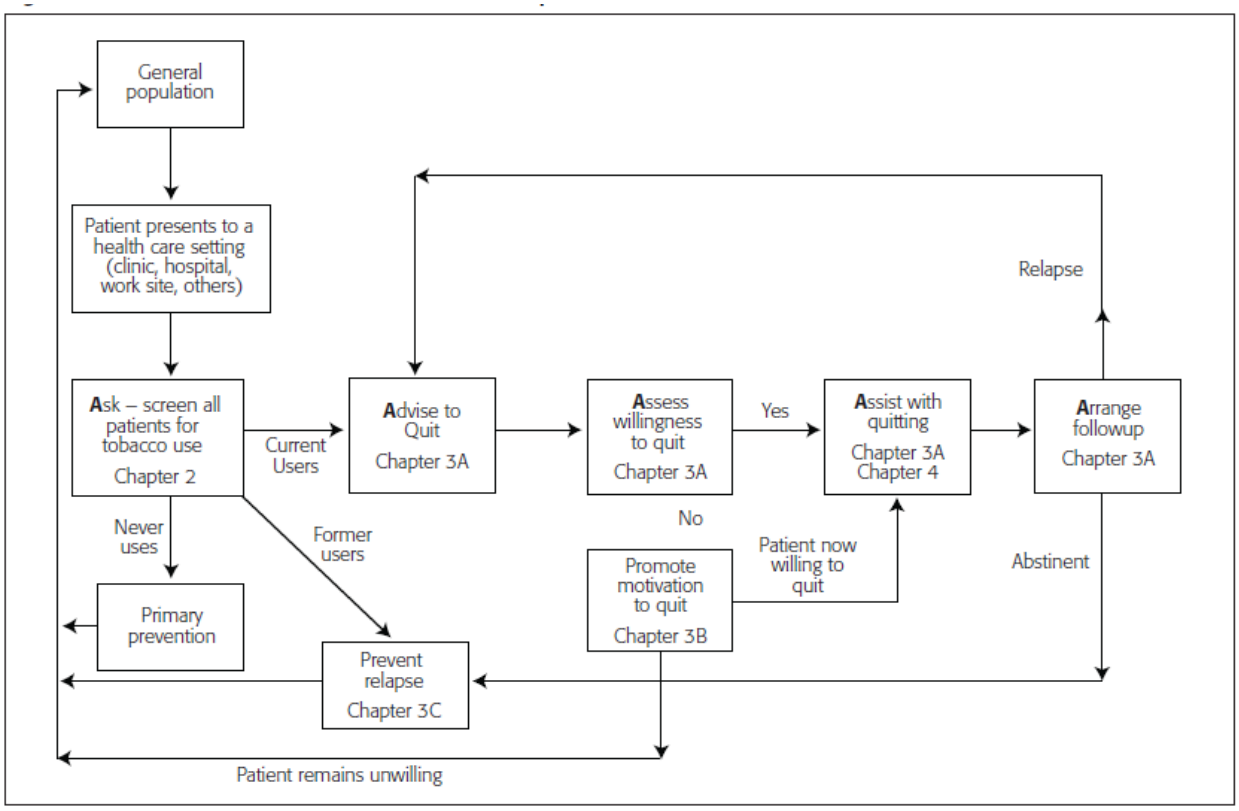
☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME PERFORMANCE MEASURE *If not a health outcome, skip to 1a.3*

1a.2. Briefly state or diagram the linkage between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.



Citation: Fiore MC, Jaen CR, Baker TB, et al. (2008) Treating tobacco use and dependence: 2008 update. Clinical practice guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) and at least one health care structure, process, intervention, or service.

This measure is intended to provide CMS and provider/provider groups that serve Medicare beneficiaries smoking prevalence data on Medicare populations served. These data will allow CMS and agencies serving CMS populations to better assess what proportion of the population being served is affected by smoking, what the variability is in smoking rates across CMS populations, and the impact provider services and CMS benefits are having on smoking rates. The Affordable Care Act (ACA) specifically contained provisions relevant to the expansion of CMS benefits related to smoking cessation for Medicare beneficiaries. Under Medicare, beneficiaries will continue to receive coverage for counseling and prescription medications for up to two quit attempts per year. In addition, Medicare copayment, coinsurance, and deductibles for cessation treatments were waived effective January 1, 2011.

Beyond the benefits CMS provides to assist in smoking cessation, CMS has also implemented a number of tobacco use-related measures for providers/provider group reporting. In 2015, 15 measures were identified in the CMS Measure Inventory. Five were currently in use in CMS programs, two were in active development, two were in proposed rule-making, and six measures were previously used in CMS programs. Nearly all measures reviewed (N=13) were process measures targeting appropriate provision of tobacco use screening and cessation services (counseling, pharmacotherapy, etc., although these process measures do not align with current CMS benefits [discussed above], which focus on changes in status). The six endorsed measures were among those most commonly used by CMS programs. For instance, NQF #0028 – the percentage of patients aged 18 years and older who were screened for tobacco use one or more times within 24 months and who received cessation counseling intervention if identified as a tobacco user – is used by five CMS programs, including Million Hearts, Physician Quality Reporting System (PQRS), Physician Compare, Value-Based Payment Modifier (VBM), Physician Feedback/Quality Resource Use Report, and the Medicare Shared Savings Program.

Use of a smoking prevalence measure is even more relevant now with the passing of the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) that creates a new Medicare quality payment program supporting two Medicare physician payment tracks. The Merit-Based Incentive Payment system (MIPS) which most Medicare physicians are expected to participate in initially, increases the relevance of value-based care by adjusting Medicare payments to providers based on four performance categories. These categories include cost of care, quality of care, clinical care improvement activities, and advancing care information. The quality of care performance category is based on quality measures that focus on population health and outcomes. While CMS does currently evaluate the percentage of the population assessed for smoking and the proportion of smokers that are provided intervention, there are currently no measures providing information on the outcome that these processes are supposed to influence – smoking prevalence. Importantly, MIPS will sunset several programs that currently use tobacco use process measures including the Physician Quality Reporting System (PQRS), Value-based Payment Modifier (VBM), and the Medicare electronic health records (EHR) Incentive Program for EPs, commonly known as Meaningful Use (MU).

Note: For health outcome performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the linkages between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

N/A

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☒ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Citation: Fiore MC, Jaen CR, Baker TB, et al. (2008) *Treating tobacco use and dependence: 2008 update*. Clinical practice guideline. Rockville, MD: U.S. Department of Health and Human Services. *Public Health Service.* May 2008.

URL: http://bphc.hrsa.gov/buckets/treatingtobacco.pdf

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1. Recommendation: All patients should be asked if they use tobacco and should have their tobacco use status documented on a regular basis. Evidence has shown that clinic screening systems, such as expanding the vital signs to include tobacco use status or the use of other reminder systems such as chart stickers or computer prompts, significantly increase rates of clinician intervention. (Strength of Evidence = A). pg. 77

2. Recommendation: All *physicians* should strongly advise every patient who smokes to quit because evidence shows that physician advice to quit smoking increases abstinence rates. (Strength of Evidence = A). pg. 82

3. Recommendation: Treatment delivered by a variety of clinician types increases abstinence rates. Therefore, all clinicians should provide smoking cessation interventions. (Strength of Evidence = A). pg. 87

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

All grades have been provided in 1a.4.2 with the individual recommendation.

All of the recommendations listed above received a Grade "A" rating. A recommendation receiving Grade "A" is defined as supported by evidence from "[m]ultiple well-designed randomized clinical trials, directly relevant to the recommendation, yield[ing] a consistent pattern of findings."

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Note: Recommendation grades are based on the strength of evidence; therefore, only one grading system was used to evaluate both recommendations and the body of evidence.

Grade	Definition
A	Multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings.
B	Some evidence from randomized clinical trials supported the recommendation, but the scientific support was not optimal. For instance, few randomized trials existed, the trials that did exist were somewhat inconsistent, or the trials were not directly relevant to the recommendation.
C	Reserved for important clinical situations in which the Panel achieved consensus on the recommendation in the absence of relevant randomized controlled trials.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Citation: Patnode CP, Henderson JT, Thompson JH, Senger CA, Fortmann SP, Whitlock EP. Behavioral Counseling and Pharmacotherapy Interventions for Tobacco Cessation in Adults, Including Pregnant Women: A Review of Reviews for

the U.S. Preventive Services Task Force. Evidence Synthesis No. 134. AHRQ Publication No. 14-05200-EF-1. Rockville, MD: Agency for Healthcare Research and Quality; 2015.

URL: http://www.ncbi.nlm.nih.gov/books/NBK321744/pdf/Bookshelf_NBK321744.pdf

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1. Recommendation: The USPSTF recommends that clinicians ask all adults about tobacco use, advise them to stop using tobacco, and provide behavioral interventions and US Food and Drug Administration (FDA)–approved pharmacotherapy for cessation to adults who use tobacco. (Grade A).

2. Recommendation: The USPSTF recommends that clinicians ask all pregnant women about tobacco use, advise them to stop using tobacco, and provide behavioral interventions for cessation to pregnant women who use tobacco. (Grade A).

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

The USPTF recommendations listed in 1a.5.2 were both provided Grade "A." A USPSTF recommendation receiving Grade "A" is defined as one where "[t]he USPSTF recommends the service [and] [t]here is high certainty that the net benefit is substantial."

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Grade	Definition	Suggestions for Practice
A	The USPSTF recommends the service. There is high certainty that the net benefit is substantial.	Offer/provide this service.
B	The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.	Offer/provide this service.
C	The USPSTF recommends against routinely providing the service. There may be considerations that support providing the service in an individual patient. There is moderate or high certainty that the net benefit is small.	Offer/provide this service only if other considerations support offering or providing the service in an individual patient.
D	The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or the harms outweigh the benefits.	Discourage the use of this service.
I	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, or poor quality, or conflicting, and the balance of benefits and harms cannot be determined.	Read the clinical considerations section of the USPSTF Recommendation Statement. If the service is offered, patients should understand the uncertainty about the balance of benefits and harms.

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Citation: *Grade Definitions*. U.S. Preventive Services Task Force. June 2016. http://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions. Accessed 24 June 2016.

URL: http://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

N/A

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Evidence reviews conducted for the Clinical Practice Guidelines and for the USPSTF Recommendations addressed assessment of tobacco use by clinicians, provision of clinical and system cessation interventions, assessment of abstinence following treatment, and use of smoking measure outcomes.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The Clinical Practice Guideline used a "Strength of Evidence" grading system to evaluate evidence supporting recommendations. This same grading system is used to assess the individual recommendations. All recommendations provided in 1a.4.2 are supported by evidence graded as "A" which is defined as "[m]ultiple well-designed randomized clinical trials, directly relevant to the recommendation, yield[ing] a consistent pattern of findings."

The USPSTF uses a "Level of Certainty" grading system to evaluate evidence supporting recommendations. Both USPSTF recommendations provided in 1a.5.2 are supported by evidence graded as "High." A "high" level of certainty is defined as "[t]he available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies."

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Clinical Practice Guideline Recommendations and USPSTF Recommendations both use grading schemes; however, grade definitions are different. The clinical practice guideline evidence grading system is the same as the recommendation grading system; therefore, grade definitions are provided in 1a.4.4.

The USPSTF Evidence Grading System is provided below:

Level of Certainty	Description
High	The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.
Moderate	The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: • The number, size, or quality of individual studies. • Inconsistency of findings across individual studies. • Limited generalizability of findings to routine primary care practice. • Lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion.
Low	The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: • The limited number or size of studies. • Important flaws in study design or methods. • Inconsistency of findings across individual studies. • Gaps in the chain of evidence. • Findings not generalizable to routine primary care practice. • Lack of information on important health outcomes. More information may allow estimation of effects on health outcomes.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 1975 to 2015

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

The Clinical Practice Guideline body of evidence included randomized placebo/comparison controlled trials of tobacco use treatment intervention randomized on the patient level from three separate systematic reviews resulting in 11 meta-analyses that support the recommendations listed in 1a.4.2.

The USPSTF Recommendations' body of evidence included 54 systematic reviews with and without meta-analysis. Twenty-two of these reviews were selected for the basis of the study's primary findings. This body of evidence support the recommendations listed in 1a.5.2.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

The evidence surrounding the recommendations from both the PHS and USPSTF regarding tobacco screening and cessation received the highest grades possible according to each of their respective criteria. The evidence from the PHS received an "A" grade, indicating that multiple well-designed randomized clinical trials, directly relevant to the recommendation, yielded a consistent pattern of findings. As for the USPSTF, there was high certainty that the net benefit of implementing each recommendation was substantial. In this context, high certainty refers to evidence that usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

A robust evidence base offers strategies to help individuals quit smoking. This evidence is reflected in both the U.S. Public Health Service Guideline on Treating Tobacco Use and Dependence, and the recently updated 2015 U.S. Preventive Services Task Force (USPSTF) recommendation for behavioral and pharmacotherapy interventions for tobacco cessation in adults. These rigorous reviews of the evidence have concluded that clinicians can substantially increase the odds that a patient will try to quit and will be successful in quitting if clinicians screen all patients for tobacco use and offer evidence-based cessation treatments to those who use tobacco.

The odds of clinicians intervening with their patients who smoke were 3.1 times higher in the presence of screening systems on the rate of smoking cessation compared to absence of such systems. However, the inclusion of screening systems alone does not improve rates of smoking cessation. Brief physician advice was shown to significantly increase long-term smoking abstinence rates. The odds of abstaining from smoking were 1.3 times higher in those who received physician advice to quit compared to those who did not receive advice.

When physicians in hospitals, general practitioner offices, or community programs referred individuals to specialist outpatient cessation treatment, patients reported more confidence to quit smoking. Individual, group, and telephone counseling have been shown to be effective in increasing quit rates, as have seven Food and Drug Administration (FDA)-

approved medications. While counseling and medication are each effective alone, the combination of counseling and medication is more effective than either component alone.

Studies on the intensity of cessation interventions show that even minimal clinician counseling (three minutes or less) yielded higher odds of abstinence compared to no contact at all. Further, a dose-response relationship is observed, with odds of abstinence increasing with intensity of interventions. This relationship is supported across many studies.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

The USPSTF reviewed behavioral intervention and pharmacotherapy studies for adverse events. In reviewing behavioral interventions, the USPSTF identified only minor adverse events associated with alternative therapies, primarily acupuncture.

The USPSTF conducted a thorough review of adverse events of pharmacotherapy on both pregnant and nonpregnant adults. In nonpregnant adults, the USPSTF found adequate evidence that the harms of nicotine replacement treatment (NRT), bupropion sustained release, or varenicline for tobacco cessation are small. The USPSTF found inadequate evidence to determine the harms of electronic nicotine delivery system (ENDS). For pregnant women, the USPSTF found inadequate evidence on the harms of NRT and no evidence on the harms of bupropion SR, varenicline, or ENDS for tobacco cessation.

Overall, the USPSTF found convincing evidence that pharmacotherapy interventions, including NRT, bupropion hydrochloride sustained-release (buproprion SR), and varenicline—with or without behavioral counseling interventions—substantially improve achievement of tobacco cessation in nonpregnant adults who smoke. However, the USPSTF found inadequate evidence on the benefits of NRT and no evidence on the benefits of bupropion SR, varenicline, or ENDS to achieve tobacco cessation in pregnant women who smoke.

NRT is a pregnancy category D medication, which means that there is evidence of fetal risk based on adverse reaction data from studies in humans. However, it has been suggested that NRT may be safer than smoking during pregnancy. Potential adverse events reported include increased rates of cesarean delivery, slightly increased diastolic blood pressure, and skin reactions to the patch. Potential adverse events reported in nonpregnant adults include higher rates of low-risk cardiovascular events, such as tachycardia. The USPSTF identified no studies on bupropion SR or varenicline pharmacotherapy during pregnancy. These drugs are both pregnancy category C, which means that animal reproduction studies have shown an adverse effect on the fetus but there are no adequate well-controlled studies in humans.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

N/A

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

N/A

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all 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

Patient_Panel_Smoking_Measure_Evidence_Attachment_01-27-17.docx

1a.1 For Maintenance of Endorsement: Is there new evidence about the measure since the last update/submission?

Please update any changes in the evidence attachment in red. Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. If there is no new evidence, no updating of the evidence information is needed.

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 PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. (Describe how and from whom their input was obtained.)
IF a COMPOSITE (e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and provide rationale for composite in question 1c.3 on the composite tab.

Health behaviors such as smoking are critical contributors to multiple poor health outcomes, and there is a large body of evidence that health behaviors are amenable to modification by practitioners in the clinical care sector. Despite declines in use since its peak in 1965, tobacco consumption, and cigarette smoking in particular, remains the single most preventable cause of disease and death in the US (HHS 2014). As of 2013, approximately one in four adults were current users of some form of tobacco, with an estimated 42.1 million (17.8%) adults smoking cigarettes (CDC 2014). Compared to the general population, the prevalence of smokers is considerably higher among Medicaid beneficiaries (37%; CMS Medicaid 2015) and lower among Medicare beneficiaries (9.9%; CDC 2014). Thus, smoking represents a high impact, high cost behavioral risk factor of broad relevance to CMS models and programs where the clinical care sector has opportunities to play an important role in screening, counseling, prescribing treatments, and facilitating referrals to community-based resources to help patients improve their health.

To date, CMS performance measures associated with smoking have focused on clinical processes related to risk assessment and appropriate follow-up activities for patients as needed. A transition to outcome measures would move the focus beyond just screening and referrals, but also on identifying and executing effective strategies for risk reduction, including both clinical interventions and clinical-community partnerships. When combined with health system-specific changes that may be needed to support efficient workflow and data feedback loops to facilitate continuous learning, current process measures for tobacco cessation may fall short of recognizing all the variation and nuance potentially needed for long-term success. Outcome measures offer an alternative that would allow for evaluation of context-specific interventions, fit well into an environment of constant adaptation and continuously emerging properties, and have applicability in other population-based initiatives targeting optimized levels of health system efficiency and cultural sensitivity. Outcome measures offer a mechanism to examine tailored approaches to operationalizing the 5 As (ask, advise, assess, assist, arrange), and to assess the impact of collaboration with external organizations or agencies representing the public health or community sectors. In shifting to outcomes, the measure focus would emphasize the end result rather than a checklist of procedural steps. Thus, the use of outcome measures could encourage the development and refinement of effective innovations best-suited and adapted to a particular context (community, health care system, patient), resulting in population health improvements via patient-centered care.

Citations:
HHS (United States Department of Health and Human Services). (2014) The health consequences of smoking—50 years of progress: A report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf. Accessed 9 September 2015.

CDC (Centers for Disease Control and Prevention). (2014) Current cigarette smoking among adults—United States, 2005-2013. MMWR 63(47):1108-1112. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6347a4.htm?s_cid=mm6347a4_e. Accessed 9 September 2015.

CMS (Centers for Medicare and Medicaid Services) Medicaid (2015) Tobacco Cessation. Accessed 9 September 2015 from: http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Benefits/Tobacco.html.

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 (4b) under Usability and Use.*

For this initial endorsement, performance scores are provided from measure testing using historical PQRS data (2014). Smoking prevalence was calculated across providers. The distribution of the measure across EPs is shown in the table below. The analyses below are at the provider level. The third column of the table represents the total number of patients that are reflected in each year of data. Lower values are better as they reflect a lower prevalence of smoking at the provider level.

Smoking Prevalence, at the Provider Level (PQRS: EHR QRDA1), 2014

Year	Providers	Patients	Mean	Std Dev	IQR	Min	10th Pctl	20th Pctl	30th Pctl	40th Pctl	Median	60th Pctl	70th Pctl	80th Pctl	90th Pctl	Max
2014	**382**	**72,478**	13.2	12.0	12.1	0.0	2.6	4.6	5.9	7.7	9.6	11.9	14.9	19.6	28.2	69.2

The table shows that there is substantial variation in smoking prevalence across providers. As expected, the variation among providers was somewhat large given the relatively small number of patients per provider and the relatively large number of providers.

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.

As noted in the rationale in 1b.1, despite declines in use since its peak in 1965, tobacco consumption, and cigarette smoking in particular, remains the single most preventable cause of disease and death in the US (HHS 2014 Health Consequences). As of 2015, an estimated 36.5 million (15.1%) adults were still smoking cigarettes (CDC 2016).

Substantial evidence indicates that clinical tobacco cessation programs can be very effective in helping smokers quit. On average, individuals who smoke make 8-11 quit attempts before successfully quitting and data have shown that the probability of success increases with each quit attempt (HHS 2001; Hughes 2003). When surveyed, nearly 70% of smokers indicated they were interested in quitting and that a doctor's advice to quit would increase their likelihood of quitting (HHS 2014 Preventing Tobacco Use). Provider intervention has been shown to significantly impact patients' success in quit attempts, and patients point to physician advice as an important motivator (Fiore 2008).

Receiving brief advice to quit is associated with a 30% increase in the number of smokers who quit, and cessation programs using either therapy, medication, or a combination of both have success rates between 10%-30% (Gorin and Heck 2004; Katz 2004; Leif Associates 2012; NCQA 2014). Because even brief interventions have been shown to be effective, providers' continued assessment and provision of education can facilitate important opportunities for patients to quit (Fiore 2000; AAMC 2007).

Strategies that have been shown to improve rates of tobacco cessation counseling and interventions in primary care settings include implementing a tobacco user identification system; providing education, resources, and feedback to promote clinician intervention; and dedicating staff to provide tobacco dependence treatment and assessing the delivery of this treatment in staff performance evaluations (Fiore 2008).

Indeed, performance on the existing PQRS tobacco screening and intervention process measure has demonstrated providers are engaging with patients on this topic. According to the 2014 PQRS Experience Report, average performance on this measures was 81.6% in 2011 and has increased over time to 88.9% in 2014.

Citations:

AAMC (Association of American Medical Colleges). (2007) Physician behavior and practice patterns related to smoking cessation. http://www.legacyforhealth.org/content/download/566/6812/file/Physicians_Study.pdf. Accessed 9 September 2015.
CDC (Centers for Disease Control and Prevention). (2016) Current Cigarette Smoking Among U.S. Adults Aged 18 Years and Older.. https://www.cdc.gov/tobacco/campaign/tips/resources/data/cigarette-smoking-in-united-states.html. Accessed 14 February 2016.
CMS (Centers for Medicare and Medicaid Services) Medicaid (2015) Tobacco Cessation. Accessed 9 September 2015 from: http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Benefits/Tobacco.html.
Fiore MC. (2000) A Clinical practice guideline for treating tobacco use dependence. JAMA 283(24):3244-3254. http://whyquit.com/guidelines/2000JuneConsensus.pdf. Accessed 9 September 2015.
Fiore MC, Jaen CR, Baker TB, et al. (2008) Treating tobacco use and dependence: 2008 update. Clinical practice guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.
Gorin SS, Heck JE. (2004) Meta-analysis of the efficacy of tobacco counseling by health care providers. Cancer Epidemiol Biomarkers Prev 13(12):2012-2022. http://cebp.aacrjournals.org/content/13/12/2012.full.pdf. Accessed 9 September 2015.
HHS (United States Department of Health Human Services). (2001) Women and smoking: A report of the Surgeon General. Rockville,MD: U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease, Prevention and Health Promotion, Office on Smoking and Health.
HHS (United States Department of Health and Human Services). (2014) The health consequences of smoking—50 years of progress: A report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf. Accessed 9 September 2015.
HHS (United States Department of Health Human Services). (2014) A Report of the Surgeon General: Preventing tobacco use among youth and young adults. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.cdc.gov/tobacco/data_statistics/sgr/2012/consumer_booklet/pdfs/consumer.pdf. Accessed 9 September 2015.
Hughes JR. (2003) Motivating and helping smokers to stop smoking. J Gen Intern Med 18:1053-1057. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1494968/. Accessed 9 September 2015.
Katz DA, Muehlenbruch DR, Brown RL, Fiore MC, Baker TB. (2004) Effectiveness of implementing the agency for healthcare research and quality smoking cessation clinical practice guideline: a randomized, controlled trial. J Natl Cancer Inst 96(8):594-603.
Leif Associates, Inc. (2012) The business case for coverage of tobacco cessation: 2012 update. http://www.ctri.wisc.edu/Employers/Actuarial.Analysis.pdf. Accessed 9 September 2015.
NCQA (National Committee for Quality Assurance). (2014) HEDIS 2015: Healthcare effectiveness data and information Set. Vol. 1, narrative. Washington (DC): National Committee for Quality Assurance (NCQA). http://www.ncqa.org/HEDISQualityMeasurement/HEDISMeasures/HEDIS2014.aspx. Accessed 9 September 2015.

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 (4b) under Usability and Use.*

For this initial endorsement, performance scores are provided from measure testing using historical PQRS data. The table below reports the smoking prevalence stratified by age, race, ethnicity, and sex.

Table: Smoking prevalence by characteristics of patients, 2014 PQRS EHR QRDA1

Variable	Group	Patients	% of Total	Smoking prevalence
Total	Total	72,478	100%	12.8%
Age	Less than 65 years	25,461	35.1%	22.7%
	65 to 69 years	12,822	17.7%	11.2%
	70 to 74 years	11,470	15.8%	8.3%
	75 to 79 years	9,549	13.2%	5.8%
	80 to 84 years	7,093	9.8%	4.6%
	85 years or older	6,083	8.4%	3.2%
Sex	Male	29,425	40.6%	14.9%
	Female	43,045	59.4%	11.3%
	Unknown	8	0.01%	25.0%
Race	AIAN	63	0.1%	25.4%
	Asian	2,063	2.9%	5.4%
	Black	3,099	4.3%	18.5%
	NHPI	24	0.03%	20.8%
	White	41,727	57.6%	11.9%
	Other	23,177	32.0%	14.2%
	Unknown	2,325	3.2%	11.8%
Ethnicity	Hispanic or Latino	1,849	2.6%	8.9%
	not Hispanic or Latino	67,868	93.6%	12.9%
	Unknown	2,761	3.8%	11.1%

AIAN=American Indian or Alaskan Native
NHPI=Native Hawaiian or Pacific Islander

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**

While a number of studies support the efficacy of provider-based tobacco cessation interventions overall, evidence of effectiveness among subgroups is limited. Despite the data suggesting that most smokers are interested in quitting, racial/ethnic minority smokers are less likely to receive advice to stop smoking, use recommended treatment to aid cessation, or succeed in achieving or maintaining abstinence (CDC 2014; HHS 2014; Fu 2008; Fu 2007; Levinson 2006; Houston 2005; Bansal 2004; Levinson 2004). Factors such as lower socioeconomic status (SES), cultural and language barriers, and reduced access to health care contribute to reduced treatment access (Lurie and Dubowitz 2007). Differences by racial/ethnic groups in smoking prevalence, patterns of use, and treatment outcomes are well documented (Fu 2008; Shiffman 2008; Fagan 2007; Okuyemi 2004; King 2004) and contribute to the need to evaluate treatment interventions within and across groups (CDC 2014; Schlundt 2007; Benowitz 2002). Others have postulated that because genetic, sociocultural, and pharmacological determinants of smoking may differ by racial/ethnic subgroup, tobacco cessation interventions may be expected to produce heterogeneous outcomes, and therefore, evaluation of effectiveness across different subgroups is warranted (Cropsey 2014; Cox 2011; Benowitz 2002).

The need for this type of research is echoed in the USPHS guidelines and USPSTF recommendations (Fiore 2008; USPSTF 2009). While the guidelines encourage universal identification of all tobacco users and the use of counseling and pharmacotherapy for treatment of smokers, the limited inclusion of racial/ethnic minority smokers within treatment research is noted. Importantly, the guidelines specifically call for additional research on effectiveness of smoking cessation interventions among racial/ethnic minority populations, efficacy of culturally adapted versus generic interventions, factors salient to treatment, and identification of barriers to enhance access and efficacy of intervention for these groups (Fiore 2008). Reports of tailored strategies are beginning to emerge in the literature. For example, tailored approaches have shown positive results for African American smokers (Matthews 2009) and, separately, for Chinese American smokers (Wu 2009). An intervention targeted to help pregnant women who smoke was able to significantly reduce smoking between the intervention group (28% quit) and the control group (9.8% quit) (Bailey 2015). Yet other examples among youth have demonstrated success via school-based initiatives (Backinger 2003; Flay 2009).

Citations:

Backinger CL, Fagen P, Matthews E, Grana R. (2003) Adolescent and young adult tobacco prevention and cessation: current status and future directions. Tob Control 12:46-53.

Bailey BA. (2015) Effectiveness of a pregnancy smoking intervention: the Tennessee Intervention for Pregnant Smokers Program. Health Educ Behav. 42(6):824-831.

Bansal MA, Cummings KM, Hyland A, Giovino GA. (2004) Stop-smoking medications: who uses them, who misuses them, and who is misinformed about them? Nicotine Tob Res 6(Suppl 3):S303-S310.

Benowitz NL. (2002) Smoking cessation trials targeted to racial and economic minority groups. JAMA 288(4):497-499.

CDC (Centers for Disease Control and Prevention). (2014) Current cigarette smoking among adults— United States, 2005-2013. MMWR 63(47):1108-1112. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6347a4.htm?s_cid=mm6347a4_e. Accessed 9 September 2015.

Cox LS, Okuyemi K, Choi WS, Ahluwalia JS. (2011) A review of tobacco use treatments in US ethnic minority populations. Am J Health Promot 25(50):S11-S30.

Cropsey KL, Leventhal AM, Stevens EN, Trent LR, Clark CB, Lahti AC, Hendricks PS. (2014) Expectancies for the effectiveness of different tobacco interventions account for racial and gender differences in motivation to quit and abstinence self-efficacy. Nicotine Tob Res 16(9):1174-1182.

Fagan P, Moolchan ET, Lawrence D, Fernander A, Ponder PK. (2007) Identifying health disparities across the tobacco continuum. Addiction 102(Suppl 2):5-29.

Fiore MC, Jaen CR, Baker TB, et al. (2008) Treating tobacco use and dependence: 2008 update. Clinical practice guideline. Rockville, MD: U.S. Department of Health and Human Services. Public Health Service. May 2008.

Flay BR. (2009) School-based smoking prevention programs with the promise of long-term effects. Tob Induc Dis 5(1):6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2669058/. Accessed 9 September 2015.

Fu SS, Burgess D, van Ryn M, Hatsukami DK, Solomon J, Joseph AM. (2007) Views on smoking cessation methods in ethnic minority communities: a qualitative investigation. Prev Med 44(3):235-240.

Fu SS, Kodl MM, Joseph AM, Hatsukami DK, Johnson EO, Breslau N, Wu B, Bierut L. (2008) Racial/ethnic disparities in the use of nicotine replacement therapy and quit ratios in lifetime smokers ages 25 to 44 years. Cancer Epidemiol Biomarkers Prev 17(7):1640-1647.

HHS (United States Department of Health and Human Services). (2014) The health consequences of smoking—50 years of progress: A report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. http://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf. Accessed 9 September 2015.

Houston TK, Scarinci IC, Person SD, Greene PG. (2005) Patient smoking cessation advice by health care providers: the role of ethnicity, socioeconomic status, and health. Am J Public Health 95(6):1056-1061.

King G, Polednak A, Bendel RB, Vilsaint MC, Nahata SB. (2004) Disparities in smoking cessation between African Americans and Whites: 1990-2000. Am J Public Health 94(11):1965-1971.

Levinson AH, Perez-Stable EJ, Espinoza P, Flores ET, Byers TE. (2004) Latinos report less use of pharmaceutical aids when trying to quit smoking. Am J Prev Med 26(2):105-111.

Levinson AH, Borrayo EA, Espinoza P, Flores ET, Perez-Stable EJ. (2006) An exploration of Latino smokers and the use of pharmaceutical aids. Am J Prev Med 31(2):167-171.

Lurie N, Dubowitz T. (2007) Health disparities and access to health. JAMA 297(10):1118-1121.

Matthews AK, Sanchez-Johnsen, King A. (2009) Development of a culturally targeted smoking cessation intervention for African American smokers. J Community Health 34(6):480-492. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3712791/. Accessed 9 September 2015.

Okuyemi KS, Ahluwalia JS, Banks R, Harris KJ, Mosier MC, Nazir N, Powell J. (2004) Differences in smoking and quitting experiences by levels of smoking among African Americans. Ethn Dis 14(1):127-133.

Shiffman S, Brockwell SE, Pillitteri JL, Gitchell JG. (2008) Use of smoking-cessation treatments in the United States. American Journal of Preventive Medicine 34: 102–111.

Schlundt DG, Niebler S, Brown A, Pichert JW, McClellan L, Carpenter D, Blockmon D, Hargreaves M. (2007) Disparities in smoking: data from the Nashville REACH 2010 project. J Ambul Care Manage 30(2):150-158.

USPSTF (United States Preventive Services Task Force). (2009) Counseling and interventions to prevent tobacco use and tobacco. Ann Intern Med 150(8):551-555.

Wu D, Ma GX, Zhou K, Liu A, Poon AN. (2009) The effect of a culturally tailored smoking cessation for Chinese American smokers. Nicotine Tob Res 11(12):1448-1457. http://www.ncbi.nlm.nih.gov/pubmed/19915080. Accessed 9 September 2015.

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. **Cross Cutting Areas** *(check all the areas that apply):*
«crosscutting_area»

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.)*
N/A

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:** Smoking_Data_Dictionary_09-21-2016.xlsx

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).
Adult patients identified as smokers.

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).
Adult patients identifed as smokers as of the last qualifying encounter during the measurement period.

Measurement Period: One year.
Qualifying encounter: Please see attached data dictionary for qualifying encounters.
Smoking Status: Adult patients must be identified as smokers within 24 months prior to the end of the measurement period (that is, during the measurement period or the year prior). This period aligns with the data collection period used for NQF 0028 (Tobacco Use: Screening and Cesssation Intervention). Please see attached data dictionary for codes that identify a smoker.

S.6. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Adult patients who had a qualifying encounter with a provider during the measurement period AND were identified as either as smokers or non-smokers within 24 months of the end of the measurement period.

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).
Measurement Period: One year.
Qualifying encounter: Please see attached data dictionary for qualifying encounters.
Smoking Status: Adult patients must be identified as either a smoker or a non-smoker within 24 months prior to the end of the measurement period (that is, during the measurement period or the year prior). This period aligns with the data collection period used for NQF 0028 (Tobacco Use: Screening and Cesssation Intervention).

S.8. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Adult patients were excluded if their smoking status (either as a smoker or a non-smoker) was missing.

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.)*
Adult patients that do not have their smoking status (either as a smoker or a non-smoker) recorded within 24 months of the end of the measurement period, for any reason, are excluded.

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 = Lower score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

To calculate 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 the 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.
4. From the patients who did not meet the numerator criteria, determine if the patient meets any criteria for exception when denominator exceptions have been specified [for this measure: missing smoking status is an exclusion criteria for the denominator]. If the patient meets any exception criteria, they should be removed from the denominator for performance calculation. Although the exception cases are removed from the denominator population for the performance calculation, the exception rate (ie, percentage with exceptions) should be calculated and reported along with performance rates to track variations in care and highlight possible areas of focus for quality improvement.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF a PRO-PM, identify whether (and how) proxy responses are allowed.

S.15 is not applicable since it does not rely on a sample of EHR QRDA1 data submitted via the PQRS data. Rather, it uses all (that is, the universe) of available EHR QRDA1 data. See above for additional details on exclusions.

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.)*

IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.16 is not applicable since it relies on data submitted to the PQRS.

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 Record (Only)

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 is collected.)*

IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.

The measure relies on data from the PQRS. PQRS is a quality reporting program that was established to encourage individual providers and group practices to report information on the quality of care to CMS. Medicare physicians, practitioners, and therapists providing covered professional services paid under or based on the Medicare Physician Fee Schedule (MPFS) are all defined as EPs under PQRS. Individual EPs, EPs in group practices participating in PQRS via Group Practice Reporting Option (GPRO), Accountable Care Organizations (ACOs) reporting for PQRS via the web interface, and Comprehensive Primary Care (CPC) practice sites are eligible to participate in PQRS. PQRS allows EPs to submit quality measure data through various reporting mechanisms. Each reporting mechanism has its own set of requirements for submission which results in differences in the data collected.

Only data submitted to PQRS from electronic health records (EHRs) can currently be used for the smoking prevalence measure. While only 5.8% of PQRS participants used EHR submission in 2014, it is one of the fastest growing reporting mechanisms and represents a potential "future state" where quality measurement is integrated with EHR systems. The EHR QRDA1 reporting method provides a standardized structure for providers (EPs or group practices) to submit data pertaining to electronic Clinical Quality Measures (eCQMs) from EHRs directly to CMS. Full specifications for eCQMs are updated and maintained by CMS in the eCQM library; these specifications provide algorithms for transforming data in EHR coding systems (SNOMED CT, HCPCS, LOINC, etc.) to meaningful and well-defined value sets, and then use value sets to calculate whether a patient is included in the numerator and denominator for a particular measure. Providers choosing to report on a PQRS measure via EHR are required to report all eligible encounters, per the measure denominator, for all payers.

PQRS reporting ended in 2016 with the publication of the MACRA Final Rule although final payment adjustments will be made through 2018. Beginning in 2017, the MIPS program will carry forward multiple elements of PQRS and other legacy reporting programs.

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 : Individual

S.21. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Behavioral Health : Outpatient, Clinician Office/Clinic, Other
If other: PQRS providers may include additional settings: Speech and Hearing Evaluation, Occupational Therapy Evaluation, and Ophthalmological Visits.

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
Patient_Panel_Smoking_Measure_Testing_Attachment_01-27-17-636211328696700206.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. (Do not remove prior testing information – include date of new information in red.)

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. (Do not remove prior testing information – include date of new information in red.)

2.3 For maintenance of endorsement
Risk adjustment: For outcome, resource use, cost, and some process measures, risk-adjustment that includes SDS factors is no longer prohibited during the SDS Trial Period (2015-2016). Please update sections 1.8, 2a2, 2b2, 2b4, and 2b6 in the Testing attachment and S.14 and S.15 in the online submission form in accordance with the requirements for the SDS Trial Period. NOTE: These sections must be updated even if SDS factors are not included in the risk-adjustment strategy. If yes, and your testing attachment does not have the additional questions for the SDS Trial please add these questions to your testing attachment:

What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care)

What were the statistical results of the analyses used to select risk factors?

Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of

between-unit effects and within-unit effects)

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** N/A
Measure Title: Patient panel adult smoking prevalence
Date of Submission: 1/27/2017
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful [16] differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☒ abstracted from electronic health record	☒ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ 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*).

All testing, except otherwise mentioned, was performed using data from eligible professionals (EPs) who reported through Electronic Health Record Quality Reporting Document Architecture Category 1 (EHR QRDA-1) in 2014.

1.3. What are the dates of the data used in testing?

Generally January 1, 2014-December 31, 2014, except where noted (e.g., some validity testing includes data from periods).

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item S.26*)**	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician

☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Data were obtained from the Medicare Physician Quality Reporting System (PQRS) EHR QRDA-1 reporting mechanism. The testing sample was determined by including all EPs that reported smoking status on at least 10 patients and reported smoking status for at least 50% of all patients; see section S.9 for more details. In 2014, the sample included 72,478 patients from 382 EPs across all regions of the United States. EPs varied in size, including 10 to 1,559 patients.

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*)

Testing was performed on adult (≥18 years old) patients in the 2014 PQRS EHR QRDA-1 sample that included an indication of smoking status. The sample included 72,478 patients from 382 EPs. The number and percentage of patients by age, sex, race, and Hispanic ethnicity are presented in the table below.

Table: Characteristics of Patients with Smoking Status in 2014 PQRS EHR QRDA-1

Variable	Group	Patients	% of Total	Smoking Prevalence
Total		72,478		12.8%
Age	Less than 65 years	25,461	35.1%	22.7%
	65 to 69 years	12,822	17.7%	11.2%
	70 to 74 years	11,470	15.8%	8,3%
	75 to 79 years	9,549	13.2%	5.8%
	80 to 84 years	7,093	9.8%	4.6%
	85 years or older	6,083	8.4%	3.2%
Sex	Male	29,425	40.6%	14.9%
	Female	43,045	59.4%	11.3%
	Unknown	8	0.01%	25.0%
Race	AIAN	63	0.1%	25.4%
	Asian	2,063	2.9%	5.4%
	Black	3,099	4.3%	18.5%
	NHPI	24	0.03%	20.8%
	White	41,727	57.6%	11.9%

	Other	23,177	32.0%	14.2%
	Unknown	2,325	3.2%	11.8%
Ethnicity	Hispanic or Latino	1,849	2.6%	8.9%
	Not Hispanic or Latino	67,868	93.6%	12.9%
	Unknown	2,761	3.8%	11.1%

AIAN=American Indian or Alaskan Native
NHPI=Native Hawaiian or Pacific Islander

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.

For most testing, the sample included 72,478 patients, with smoking status reported from 382 EPs in 2014.

For validity testing, the sample comprised a subset of 106 EPs. First, we restricted to 111 EPs who met the minimum criteria for reporting smoking status, not only in 2014, but also in 2015. Then, the test was further limited to 106 EPs that had at least a 50% reporting rate on the PQRS measure "tobacco use screening and intervention" (NQF #0028) in 2014. The sample of 106 EPs included 13,914 patients in 2014 and 17,742 patients in 2015; smoking prevalence was 10.8% in 2014 and 11.9% in 2015.

For exclusion testing, the sample included all patients (not just those with smoking status reported) from the 382 EPs. The sample included 104,395 patients from the 382 EPs in 2014. The exclusion testing assesses the potential impact of excluding patients without smoking status reported among the 382 EPs.

For missing data and minimizing bias testing, the sample included EPs with at least one patient with smoking status reported (not just the 382 meeting the minimum reporting criteria). This sample included 142,304 patients from 1,718 EPs in 2014. The missing data testing assesses the impact of restricting the sample to providers with at least 10 patients and at least 50% of patients with smoking status reported.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Patient-level SDS variables in PQRS EHR QRDA-1 are limited to patient characteristics, including age, sex, race, and ethnicity (see section 1.6). Patient and provider zip codes can be used to obtain proxy and community characteristics.

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*)

Performance measure score

Measures of homogeneity among EPs (referred to as a "unit" hereafter) were used to estimate reliability of measure scores. Reliability is the proportion of the measure variability that is attributable to the between-units variance, that is, between EPs. Reliability measures the ratio of signal to noise: the proportion of the variability in smoking prevalence that can be explained by real differences in EPs (signal), as opposed to random variation among people (noise).

A reliability value near zero indicates most of the variation observed between units is driven by random noise (within-unit variance), and hence the measure is not necessarily reflective of the unit. A reliability value near one indicates most of the variation between units is due to the differences between units, as opposed to random variation among people. A reliability value of at least 0.7 is considered sufficient to distinguish performance between two units (Adams 2009).

Reliability was estimated from a beta-binomial model. The beta-binomial model assumes the unit's score is a binomial random variable conditional on the unit's true value that comes from the beta distribution. Thus, the beta-binomial is very apt for simple pass/fail proportion measures (Adams 2009). Alpha and beta estimates from the beta-binomial model were used to calculate the unit-to-unit variation. $\left(\sigma^2_{provider-provider}\right)$

Reliability = Between Unit Variance / [Between Unit Variance + Within Unit Variance]

$$\text{Reliability} = \frac{\sigma^2_{provider-provider}}{\sigma^2_{provider-provider} + \sigma^2_{binomial}}$$

$$\text{Reliability} = \frac{\sigma^2_{provider-provider}}{\sigma^2_{provider-provider} + \frac{p(1-p)}{n}}$$ where p is the true proportion of the unit.

For the reliability calculation, p is estimated as $\hat{p} = \frac{\#\ of\ smokers}{\#\ of\ patients}$

Using these values, a reliability value was calculated for each unit. The primary result of interest is the mean reliability value across all units; additional results include reliability among subgroups (small units) and variation in reliability (e.g., interquartile range [IQR]).

Reliability was calculated overall and also for different size categories to inform decisions about minimum size for reliable reporting.

Smoking prevalence varies considerably by region and between urban and rural areas. A reliable measure should demonstrate variation between providers within regions, not rely on regional differences in smoking prevalence. Thus, EP zip codes were used to calculate reliability by census regions and rural-urban commuting area (RUCA).

Citation:

Adams, John L (2009). The Reliability of Provider Profiling. Available on-line at http://www.rand.org/pubs/technical_reports/TR653.html. Accessed 9/1/16.

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*)

Performance measure score reliability

The mean and median reliability among EPs was 0.899 and 0.946, respectively. Half of the providers had a reliability ranging from 0.885 (25th percentile) to 0.976 (75th percentile). Reliability was high regardless of the number of patients

treated by the EP, although there was a modest increase in reliability when testing was restricted to larger providers. Reliability remained high when stratifying by census regions and urban/rural status. Reliability varied somewhat within these categories, with the lowest observed in the Northeast (mean=0.819) and the highest observed in the Midwest (mean=0.903). Reliability was slightly higher in urban areas than in rural areas.

Table: Smoking Prevalence Reliability for Providers Reporting in 2014 PQRS EHR QRDA-1

Group	Eligible Professionals	Patients	Reliability *Mean, Median [25th, 75th]*
Overall	382	72,478	0.899, 0.946 [0.885, 0.976]
Minimum Provider Size			
≥ 20 patients	364	72,197	0.912, 0.946 [0.890, 0.975]
≥ 30 patients	344	71,722	0.922, 0.946 [0.897, 0.975]
≥ 40 patients	328	71,186	0.926, 0.947 [0.904, 0.975]
≥ 50 patients	314	70,568	0.931, 0.948 [0.910, 0.976]
≥ 100 patients	242	65,341	0.948, 0.957 [0.930, 0.979]
Census Region			
Northeast	54	8,563	0.819, 0.872 [0.758, 0.932]
Midwest	56	9,086	0.903, 0.941 [0.877, 0.974]
South	189	36,196	0.889, 0.927 [0.870, 0.966]
West	83	18,633	0.885, 0.938 [0.863, 0.972]
RUCA			
Urban	351	64,986	0.898, 0.947 [0.885, 0.978]
Rural	31	7,492	0.878, 0.910 [0.825, 0.940]

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?*)

Performance measure score reliability

The mean reliability of 0.899 indicated that 89.9% of the variation in smoking prevalence can be attributed to differences among EPs (signal) and 10.1% to differences among patients within EPs (noise). This was considered to be high reliability (>0.7; Adams 2009), and it showed that most of the variation observed in the measure score is due to the differences among EPs as opposed to random variation among patients within EPs. Reliability levels remained high when stratifying by region and urban/rural status, indicating that meaningful differences exist among EPs within regions, and that differences in smoking prevalence at the EP level are not entirely due to regional differences. While the number of patients included for an EP varied considerably, mean reliability and the IQR of reliability across EPs was high, regardless of any restrictions on minimum number of patients.

Citations:

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas.1960;20:37–46.
Landis, J.R.; Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics. 33 (1): 159–174.

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd edition).

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Face validity

Face validity of the measure score as an indicator of population health was assessed by the Health Behaviors Technical Expert Panel (TEP) in accordance with guidance from the Centers for Medicare and Medicaid Services (CMS) Measures Management System (MMS) Blueprint and prior practice related to the development of NQF measure #0028 (tobacco use process measure).

The panel consisted of nine members whose subject matter expertise included:

- Tobacco-related interventions and outcomes
- Population health measurement
- Health information systems
- Performance measurement
- Quality improvement
- Health care delivery perspective
- Purchaser perspective
- Person/family perspective

Our expert panel roster follows below:

Name, Credentials	Organizational Affiliation, City, State	Conflict of Interest Disclosure
Amy Aronsky, DO, FCCP, FAASM	CareCentrix, Hartford, CT	None
Steven Bernstein, AB, MA, MD	Yale University School of Medicine, New Haven, CT	None
Kevin Fontaine, MA, PhD	Department of Health Behavior, University of Alabama, Birmingham School of Public Health, Birmingham, AL	None
Aaron Garman, MD	Coal County, Community Health Center, Beulah, ND	None
Matthew Haemer, MD, MPH	Department of Pediatrics, Section of Nutrition, University of Colorado School of Medicine, Aurora, CO	None
Trina Histon, PhD	Obesity Prevention and Treatment, Behavior Change, Kaiser Permanente, Oakland, CA	None
Marc Manley, MD, MPH *(TEP Chair)*	Shoreview, MN	None

Name, Credentials	Organizational Affiliation, City, State	Conflict of Interest Disclosure
Marjorie Mitchell, MA	Michigan Universal Health Care Access Network (MICHUHCAN), Farmington, MI	None
Kenneth Warner, PhD	University of Michigan School of Public Health, Ann Arbor, MI	None

Once the measure was fully specified following iterative testing and refinement, the TEP was asked to vote electronically on the following statement:

"The scores obtained from the measure as specified will provide an accurate reflection of population health and can be used to distinguish good and poor population health".

1=Strongly Disagree; 2=Disagree; 3=Neither Agree nor Disagree; 4=Agree; 5=Strongly Agree

Empirical validity testing

Empirical validity testing was performed for the patient panel level by relating the candidate measure to a related process measure where a theoretical relationship can be tested using statistical regression modeling. Validity analyses were conducted with the PQRS EHR QRDA-1 data using the NQF-endorsed tobacco use screening and intervention measure (NQF #0028). The data sample consisted of 106 EPs with sufficient smoking prevalence data in both 2014 and 2015 and also had sufficient data (at least a 50% reporting rate) on tobacco use screening and intervention in 2014 (NQF #0028).

The hypothesized relationship was that EPs who more often engage in more tobacco use screening and intervention in 2014 would be the EPs that have lower smoking prevalence in 2015. The tobacco use screening and intervention measure was calculated at the EP level for 2014, and its correlation with smoking prevalence in 2015 was tested. Additionally, to evaluate the pattern and the relationship, tobacco use screening and intervention measure was categorized into tertiles, and mean smoking prevalence for 2014 was calculated for the tertiles. To test the hypothesis further, a simple regression model was fitted to evaluate the effect of 2014 tobacco screening and intervention (independent variable) on 2015 smoking prevalence (dependent variable), controlling for 2014 smoking prevalence (independent variable).

This observational analysis has the typical limitation that statistical correlation does not prove causation. Another limitation is the existing screening and intervention measure is a tobacco use measure, whereas the proposed measure is restricted to one type of tobacco use (smoking).

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Face validity

Six out of eight TEP members agreed or strongly agreed with the statement "the scores obtained from the measure as specified will provide an accurate reflection of population health and can be used to distinguish good and poor population health." No TEP members disagreed with this statement. Two TEP members responded as "neither agree nor disagree". One TEP member did not vote for the face validity question.

Empirical validity testing

The correlation between tobacco use screening and intervention in 2014 and smoking prevalence in 2015 for 106 EPs was in the hypothesized direction, considerably strong, and statistically significant (r= -0.29; p=0.002). The table below shows smoking prevalence in 2015 by tertiles of tobacco use screening and intervention in 2014. As 2014 tobacco use screening and intervention increases, 2015 smoking prevalence decreases.

Table: 2015 Smoking Prevalence by 2014 Tobacco Use Screening and Intervention

2014 Tobacco Use Screening and Intervention	Eligible Professionals	Mean 2015 Smoking Prevalence (%)
38.2% to 58.7% of EP's patients	36	15.3
58.7% to 75.9% of EP's patients	35	12.9
75.9% to 96.0% of EP's patients	35	10.4

Linear regression model results in the table below show smoking prevalence in 2015 decreased by an estimated two percentage points for each ten percentage point increase in tobacco use screening and intervention rate (slope estimate -0.199, p=0.002).

Table: Regression Model Results

Model Parameter	Estimate [Standard Error]	P-value
Intercept	26.4 [4.4]	<0.0001
Slope	-0.199 [0.064]	0.002

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Face validity

A majority of TEP members agreed the measures have face validity for population health; no TEP member disagreed (two neither agreed nor disagreed).

Empirical validity testing

Statistical testing supports the hypothesized negative relationship between the proposed smoking prevalence measure and a related process measure of screening and intervention, which can be controlled by the provider. Results showed that more tobacco use screening and intervention in a year is associated with lower smoking prevalence in the subsequent year. Results also indicate EPs that perform well on tobacco use screening and intervention also perform well on smoking prevalence.

2b3. EXCLUSIONS ANALYSIS

☐ **no exclusions — *skip to section 2b4***

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

All patients from the 382 EPs were included for measure exclusion testing, not just patients with smoking status reported. The sample includes 104,395 patients with encounters with the 382 EPs in 2014. The denominator for the measure includes only those who had a smoking status reported (smoker, non-smoker). Patients were excluded if they

were <18 years old, had limited life expectancy, had a medical reason not to be screened, or had smoking status missing. The reasons for exclusion and their corresponding frequencies were assessed in this section.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

For all 382 EPs included in the testing sample for 2014, the table below shows the reasons why patients were excluded from the total population (n=104,395 patients).

Table: Smoking Prevalence Exclusions (PQRS: EHR QRDA-1), 2014

Total Patients	Patients Included	Patients Excluded for Specified Reasons*	Patients Excluded for Missing Smoking Status
104,395	72,478	4,207 (4.0%)	27,710 (26.5%)

* Patients were excluded for age <18, limited life expectancy, or medical reasons.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Patients were excluded from the initial population if they were children (age <18), had evidence of limited life expectancy, or for medical reasons. This consisted of 4,207 (4.0%) of the initial population of patients among the 382 providers. The primary cause of exclusion is missing data, with 27,710 (26.5%) patients excluded because they did not have smoking status reported by the EP. This level of missingness could alter smoking prevalence measures: The estimated raw smoking prevalence (# smokers/ # patients) based on the included sample is (9,252/72,478) = 12.8%. There were 100,188 patients (104,395-4,207 = 100,188) that were not excluded for limited life expectancy, or for medical reasons. Therefore, if all patients with no smoking status (27,710 patients) reported were non-smokers, then the raw smoking prevalence could be as low as (9,252/100,188) = 9.2%, and if all of these patients were smokers, then the raw smoking prevalence could be as high as ((9,252+27,710)/100,188) = 36.9%. While there was no incentive for reporting non-smokers versus smokers in the testing data, reporting standards will be developed upon implementation to ensure valid measurement.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.
This measure is intended to support population health improvement efforts at actionable levels, such as providers. In

accordance with the goal of population health improvement, the proposed smoking prevalence measure will not be risk-adjusted. Population health improvement uses of this measure will include identifying EPs with populations of greatest health risk and opportunity for behavior change interventions, including health care system interventions and those traditionally administered outside the health care system (e.g., social services, quit lines). The TEP encouraged and strongly recommended reporting of the unadjusted measure for the purpose of population health improvement.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
N/A

2b4.4a. What were the statistical results of the analyses used to select risk factors?
N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
N/A

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
N/A

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
N/A

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
N/A

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
N/A

2b4.9. Results of Risk Stratification Analysis:
N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
N/A

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Information on distributions, as well as measures of central tendency, variability, and dispersion, are presented in this section to demonstrate meaningful differences in performance. The distribution of the measure was assessed by reporting mean, median, mode, minimum, maximum, standard deviation, range, and IQR for different thresholds of number of patients.

Differences in measure performance were then evaluated for EPs (referred to as a "unit" hereafter) using patient-level data. Each unit's smoking prevalence was compared with the overall national proportion (average of smoking prevalence measure across all EPs). The differences in proportions were compared using two-sided one sample binomial proportion tests, using a conservative alpha of 0.01. In order to reduce type one error where differences are flagged as statistically significant but true values are not different, p-values generated by each test were further adjusted for multiple comparisons using Hochberg's step-up method. The direction and magnitude of each unit's smoking prevalence with respect to the overall national proportion was represented with a Z-score. The expectation was that a unit's smoking prevalence will be similar to the overall national proportion, but some units will have lower (or higher) proportions, which are considered better (or worse) than the overall national proportion. The number and percent of units were summarized into one of three categories based on the following criteria:

1. *Better than expected:*
   Smoking prevalence of 0% or Z-score <0, with a significantly lower smoking prevalence than the overall national proportion
2. *As expected:*
   Z-score = 0 or no significant difference in smoking prevalence
3. *Worse than expected:*
   Smoking prevalence of 100% or Z-score >0, with a significantly higher smoking prevalence than the overall national proportion

The categories were further stratified by the size of the units. The average smoking prevalence within each category was reported for the entire distribution and by size category. This analysis demonstrated the ability to identify practical and significant differences in performances across units.

Additionally, power analyses were performed where the minimum sample size required detecting 5%, 10%, 15%, and 20% differences in smoking prevalence (effect size) between a unit and the national average. These analyses also varied the statistical power using 50, 60, 70, 80, and 90. The number of units in the distribution that met the minimum sample size requirement in each scenario was also reported.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

The first table below shows the variation in smoking prevalence among 382 EPs in 2014. Half of the EPs have a smoking prevalence ranging from 5.3% to 17.4%, an IQR of 12.1, and mean smoking prevalence of 13.2%. Classification of EPs as better than, no different from, or worse than the national average is shown in the second table, and results of the power analyses are described in the third table.

Table: Distribution of Smoking Prevalence, by EP (PQRS: EHR QRDA-1)

Year	*EPs*	*Mean*	*Median*	*Mode*	*Std Dev*	*25th Pctl*	*75th Pctl*	*IQR*	*Min*	*Max*	*Range*
2014	382	13.2	9.6	0.0	12.0	5.3	17.4	12.1	0.0	69.2	69.2

Table: Classification of EPs for Smoking Prevalence

Categories	*EPs*	*Percent of EPs*	*Smoking Prevalence (mean)*
Better than expected	**58**	**15.2**	**2.8**
<50 patients	11	16.2	0
50-99 patients	6	8.3	1.3
100-249 patients	18	11.3	2.7
≥250 patients	23	27.7	4.5
As expected	**274**	**71.7**	**11.2**
<50 patients	50	73.5	15.2
50-99 patients	57	79.2	10.5
100-249 patients	121	76.1	9.9
≥250 patients	47	56.6	11.5
Worse than expected	**50**	**13.1**	**36.0**
<50 patients	7	10.3	43.8
50-99 patients	9	12.5	35.0
100-249 patients	20	12.6	37.4
≥250 patients	13	15.7	31.5

Table: Minimum Sample Size and Number of EPs, by Selected Statistical Power Levels

	5% Difference		***10% Difference***		***15% Difference***		***20% Difference***	
Power	*Min Sample Size*	*# of EPs Meeting Min*	*Min Sample Size*	*# of EPs Meeting Min*	*Min Sample Size*	*# of EPs Meeting Min*	*Min Sample Size*	*# of EPs Meeting Min*
50	305	58	77	274	34	335	19	368
60	377	45	96	246	44	322	25	353
70	462	34	120	213	55	306	32	339
80	573	25	151	166	70	281	41	325
90	747	8	200	116	94	247	55	306

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

Measure testing demonstrated that there was quantifiable variation among EPs. Results showed a 12.1 percentage point gap in performance between EPs at the 25th and 75th percentiles in 2014. EPs at the 75th percentile have a smoking prevalence that is almost three times that of EPs at the 25th percentile.

There were 58 EPs (15.2% of the total 382 EPs) who were better than expected ($p<0.01$), with a mean smoking prevalence of 2.8%, and 50 providers (13.1% of 382) who were worse than expected with a mean smoking prevalence of 36%. There were 274 EPs (71.7%) who were classified as no different than the overall national average. The measure was also able to distinguish statistically significant differences ($p<0.01$) in performance among providers in similar size categories.

Sample size calculations for one-sample proportion tests can be used as a guide for setting minimum sample size thresholds for plans when evaluating measure performance in quality improvement and quality performance reporting settings. As shown above, a 10% difference in smoking prevalence from the national average of 13.6% could be detected for EPs with a minimum of 120 patients, with a statistical power of 0.70 and p-value threshold (alpha) of 0.01. In 2014, 213 (55.8%) of 382 EPs met the minimum enrollee sample size requirements to detect this size of a difference. Smaller EPs would still be able to detect larger differences at the same level of power.

All the above analyses demonstrated the ability to detect meaningful differences in measure performance rates across EPs.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note*: This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms*

for the test conducted)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

As explained in section 1.7, the initial denominator consisted of 1,718 EPs who had at least one patient. In order to minimize bias, the provider count was restricted to having at least 10 patients and reporting rate ≥50%. The reporting rate is the number of patients who had smoking status reported divided by the total number of patients for the EP. Sensitivity analyses were performed to show the smoking prevalence rate by these two criteria.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

The mean smoking prevalence in 2014 among providers with a smoking status reporting rate of at least 50% and a 10 patient minimum was 13.2%, which is very close to the 2014 estimate from the Medicare Advantage Health Outcome Survey (mean=13.6%). In comparison, the subset of providers without any restrictions had an average smoking prevalence of 16.8%, which is 3.6 percentage points higher. This distribution is also more skewed, with higher smoking prevalence rates of 22.4% and 37.5% at the 75th and 90th percentiles, 0% smoking at the 10th percentile, and 100% smoking prevalence at the maximum.

Table: Smoking Prevalence for EPs, With or Without Restrictions, PQRS: EHR QRDA-1 2014

Restrictions	N	mean	min	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	max
Smoking status reporting <50% **and/or** <10 patients for provider	1,336	16.8	0	0	5.6	12.5	22.4	37.5	100
Smoking status reporting ≥50% **and** ≥10 patients for provider	382	13.2	0	2.6	5.3	9.6	17.4	28.2	69.2

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

There is no evidence to suggest that under-reporting is associated with better performance (lower smoking prevalence); however, there was some evidence to suggest that EPs who do not screen enough patients for smoking status, or who under-report, or who have very few patients, were subject to biased/less reliable estimates of smoking prevalence. If no restrictions are used, smoking prevalence may be over-estimated, and the sample may contain EPs with extreme smoking prevalence as high as 100%. The minimum smoking status reporting rate of 50% and 10-patient minimum

sample size restrictions were used in testing to obtain reasonable and reliable estimates. Thus, minimum reporting criteria may need to be implemented in measure rollout.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields** (*i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields*) Update this field for **maintenance of endorsement**.

ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.** For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.**

Attachment: Smoking_eMeasure_Feasiblity_Scorecard_09-21-2016.xlsx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PRO data (patients, service recipients, respondents) and those whose performance is being measured.

This measure was found to be reliable and feasible for implementation. The measure does not pose any new cost, burden, or other implications for providers who report these value codes. To fully implement this measure, CMS may need to establish minimum data reporting standards.

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)*.

For the provider panel specification of this measure, the following code sets are used: CPT, SNOMEDCT, and HCPCS. Use of these code sets is in accordance with the National Library of Medicine Value Set Authority Center (VSAC) free Unified Medical Language

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Specific Plan for Use	Current Use (for current use provide URL)
Public Reporting Payment Program Quality Improvement (external benchmarking to organizations) Quality Improvement (Internal to the specific organization)	

4a.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

N/A - new measure

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

N/A - new measure

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

CMS continues to finalize plans for the proposed measure. At this time, CMS intends to use the proposed measure for public reporting within six years (in accordance with NQF guidelines). CMS is considering use of the proposed measure for alternative

payment models including the Medicare Shared Savings Program and CMS Innovation Center models. Prior to public reporting, the measure will be submitted for review under applicable processes [e.g., pre-rulemaking through the Measures Under Consideration (MUC) List for review by the NQF's Measure Application Partnership (MAP)].

Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b. **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.
N/A - new measure

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.**
N/A - new measure

4c.2. **Please explain any unexpected benefits from implementation of this measure.**
N/A - new measure

4d1.1. **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 - new measure

4d1.2. **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 - new measure

4d2.1. **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 - new measure

4d2.2. **Summarize the feedback obtained from those being measured.**
N/A - new measure

4d2.3. **Summarize the feedback obtained from other users**
N/A - new measure

4d.3. **Describe how the feedback described in 4d.2 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 - new 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)**

2020 : Adult Current Smoking Prevalence

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications harmonized to the extent possible?

Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

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.)

The measure NQF 2020: Adult Current Smoking Prevalence is considered related, but not competing, because it has the same measure focus, but it does not have the same target population. The NQF 2020: Adult Current Smoking Prevalence measure targets the US population at the national and state level while the proposed Patient Panel Adult Smoking Prevalence measure targets persons attributable to a provider.

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:** Katherine, Sapra, katherine.sapra@cms.hhs.gov, 410-786-8969-
Co.3 **Measure Developer if different from Measure Steward:** Centers for Medicare & Medicaid Services
Co.4 **Point of Contact:** Katherine, Sapra, katherine.sapra@cms.hhs.gov, 410-786-8969-

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.
Amy Aronsky, DO, FCCP, FAASM (CareCentrix)
Steven Bernstein, AB, MA, MD (Yale University School of Medicine)
Kevin Fontaine, MA, PhD (Department of Health Behavior, University of Alabama, Birmingham School of Public Health)
Aaron Garman, MD (Coal County, Community Health Center)
Matthew Haemer, MD, MPH (Department of Pediatrics, Section of Nutrition, University of Colorado School of Medicine)
Trina Histon, PhD (Obesity Prevention and Treatment, Behavior Change, Kaiser Permanente)
Marc Manley, MD, MPH (Shoreview, MN)
Marjorie Mitchell, MA (Michigan Universal Health Care Access Network (MICHUHCAN))
Ken Warner, PhD, M. Phil (University of Michigan School of Public Health)

A Health Behaviors Technical Expert Panel (TEP) was convened between October 2015 and February 2016 for a total of three substantive meetings (TEP members and their associated organizations are listed above). The role of the TEP in measure development was to guide, review, and evaluate research conducted by the developer to assess candidate measures according to the NQF measure evaluation criteria.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**