NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Click to go to the link. ALT + LEFT ARROW to return

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 2152

Measure Title: Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling

Measure Steward: PCPI Foundation

Brief Description of Measure: Percentage of patients aged 18 years and older who were screened for unhealthy alcohol use using a systematic screening method at least once within the last 24 months AND who received brief counseling if identified as an unhealthy alcohol user

Developer Rationale: This measure is intended to promote unhealthy alcohol use screening and brief counseling which have been shown to be effective in reducing alcohol consumption, particularly in primary care settings.

Numerator Statement: Patients who were screened for unhealthy alcohol use using a systematic screening method at least once within the last 24 months AND who received brief counseling if identified as an unhealthy alcohol 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 unhealthy alcohol use (eg, limited life expectancy, other medical reasons)

Measure Type: Process

Data Source: Registry Data

Level of Analysis: Clinician : Group/Practice, Clinician : Individual

Original Endorsement Date: Mar 04, 2014 **Most Recent Endorsement Date:** Mar 04, 2014

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. evidence

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

1a. Evidence. The evidence requirements for a *structure, process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured. For measures derived from patient report, evidence also should demonstrate that the target population values the measured process or structure and finds it meaningful.

The developer provides the following evidence for this measure:

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

Evidence Summary or **Summary of prior review in 2014**

The developer included the following systematic review for evidence:

- The developers cited and summarized a single meta-analysis of 23 randomized control studies (summarized across 38 scientific publications). That analysis was published in the Annals of Internal Medicine in 2012, and ultimately notes that the evidence for the screening of adults for alcohol use disorders carries a B grade (i.e., moderate evidence) per AHRQ and US Preventative Taskforce Criteria. That evidence thus moderately supports the connection between screening and favorable outcomes that include: marked reductions in alcohol use and heavy drinking.
- The measure was reviewed by NQF in 2014 using the same body of evidence.

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 notes the current guideline recommendation is under review by the USPSTF. However, the final date for the posting of the updated guideline has not been disclosed. No changes were made to the draft recommendation statement that would affect this measure.

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

Updates:

Exception to evidence

N/A

Questions for the Committee:

- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review of 03/04/14. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

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

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

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer provides a rationale for the measure based on a summary of evidence from on meta-analysis from 2012. The measure intends to advance primary care located screening for unhealthy alcohol and brief counseling when such unhealthy activity is detected.
- The most recent data publically reported included in CMS' Physician Quality Reporting System (PQRS) program from 2012 through 2015 are as follows (percentage of those receiving screening for unhealthy alcohol use):
  - 2012: 74.5%
  - 2013: 75.5%
  - 2014: 66.2%
  - 2015: 74.0%
  - 2016: 68.7%* *2016 rate contrasts above rates because it requires brief intervention completion for an observation to be included in the numerator (i.e., is the entire measure) and above rates pertain only to screening [additional information provided by the developer on 12.14.18].
- Additional data analysis provided for 2015 PQRS data are as follows**:
  - Performance 10th Percentile: 19.80
  - Performance 25th Percentile: 56.60
  - Performance 50th Percentile: 84.62
  - Performance 75th Percentile: 100.00
  - Performance 100th Percentile: 100.00
  - Performance Interquartile Range: 43.41

  **Notes: n not provided, and 2015 data focuses only on the screening for unhealthy alcohol use component of the measure and not the brief counseling component.
- A current version of the measure is included in the Merit–based Incentive Payment System (MIPS). Current data are not available.

Disparities

- Developer does not have access to disparities data for this measure. In lieu of disparities data for this measure the developer provided a summary of data from the scientific literature. This literature demonstrates racial and ethnic differences in the risk for alcohol use disorders and racial, ethnic and educational differences in the prevalence of screening for unhealthy alcohol use and binge drinking.

Questions for the Committee:

- Does the data above demonstrate a sufficient performance gap specific to the measure (i.e., screening and brief intervention)?
  - Should brief intervention rates, per se, be explored more closely in the gap analysis?
- Is the current rate of 68.7%, in and of itself, persuasive as a measure of "gap" for this indicator (i.e., demonstrating clear need for population improvement)?

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

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

1a. Evidence

Comments:

**New studies were identified by the developer.
**Evidence is moderate that this results in long-term alcohol misuse.
**Process measure of delivery of care; but clearly linked to better outcomes; evidence base has only continued to deepen.
**This is a maintenance measure. I agree with the pre-evaluation comments. The evidence is moderate.
**I am unclear as to the evidence that a screening every 24 months is worthwhile. seems annual should be the low bar.
**There is much variability in alcohol screening among primary care providers; this process measure addresses a significant problem across the U.S.; an additional 8 studies are cited that addresses the low rates of alcohol misuse sreening and brief counseling.
**Maintenance measure, new data submitted covering 2015. reviewed again by USPSTF with a release date of November 2018,again rating with moderate and B Grade.
**Process measure--B USPSTF rec.

1b. Performance Gap

Comments:

**The performance gap continues to be significant; disparities are identified
**Performance data was included. 75% for a high overall (and for three years) is decent.
**Significant gaps in care still present (in essentially all populations studied); the developer doesn't collect disparities data because of burden but other information suggests that disparities are likely. given overall low performance, gaps in specific measurement of disparities here seem less concerning than they otherwise might be.
**It remains valuable to include this as a quality measure to continue to encourage primary care providers to screen for alcohol misuse and perform a brief intervention when needed. There is evidence that prevalence of screening continues ot have room for improvement.
**Yes there is a performance gap.
**There is a definite gap in care to warrent this measure; disparities data for analysis on screening is currently not available although higher rates of high risk drinking have been documented for Native Americans and Hispanics.
**8 new studies demonstrating gap and 2015 measure data on screening gap is same as 2012.
**Moderate gap, but no performance improvement in 4 years. No disparities data.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications **and** Testing

2b. Validity: Testing; Exclusions; Risk-Adjustment; Meaningful Differences; Comparability Missing Data

2c. For composite measures: empirical analysis support composite approach

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the

measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For maintenance measures – less emphasis if no new testing data provided.

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

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

Evaluators: NQF Staff

Review A

Evaluation of Reliability and Validity:

NQF staff reviewed this measure. A summary of the measure is provided below:

Reliability

- The developer conducted measure score level reliability testing.
- Using 2016 PQRS registry data, the developer used a beta-binominal model to assess the signal-to-noise ratio.
- Adams' R at the provider (physician) level was = 0.99
- Testing done at the group practice level where such practices could be individual clinicians or groups where multiple clinicians are nested together as a singular entity. The developers thus request the measure be endorsed as appropriate for individual or group clinician analyses.
- Previous reliability, ie., Kappas on numerator and denominator, were not considered for this current evaluation, because this reliability was done by comparing EHR to abstracted chart results whereas the current submission only focuses on registry data (i.e., neither EHR or charts)

Validity

- Validity testing was performed for the measure score.
- Conducted correlation analysis with two measures Preventive Care and Screening: Screening for High Blood Pressure and Follow-up Documented (PQRS #317) and Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan measure (PQRS #134)
- Results: Developer found a positive correlation between the measures:
    - **PQRS #317**

      Coefficient of correlation = 0.29

      P-value < 0.00001
    - **PQRS #134**

      Coefficient of correlation = 0.61

      P-value < 0.00001

Questions for the Committee regarding reliability:

- Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?
- The staff is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

- Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?
- The staff is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

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

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

Evaluation A: Scientific Acceptability

Measure Number: 2152

Measure Title: Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling

Type of measure:

☒ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☐ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☐ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**

☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☒ **Registry Data**

☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

☒ **Clinician: Group/Practice** ☒ **Clinician: Individual** ☐ **Facility** ☐ **Health Plan**

☐ **Population: Community, County or City** ☐ **Population: Regional and State**

☐ **Integrated Delivery System** ☐ **Other**

Measure is:

☐ **New** ☒ **Previously endorsed (**NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

RELIABILITY: SPECIFICATIONS

1. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented?** ☒ **Yes** ☐ **No**

   Submission document: "MIF_xxxx" document, items S.1-S.22

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

2. **Briefly summarize any concerns about the measure specifications.**
   - Given that testing was done with group practice level data (including practices with only one clinician), is it appropriate to consider this measure for both group and individual clinician performance assessment?

RELIABILITY: TESTING

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level** ☒ **Measure score** ☐ **Data element** ☐ **Neither**
4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**
   ☒ **Yes** ☐ **No**
5. If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ **Yes** ☐ **No**
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   - The developer used a beta-binominal model to assess the signal-to-noise ratio.
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   - The reliability of the measure score was assessed using 2016 PQRS registry data.
   - Results of reliability testing was 0.99 using Adams' R calculation, a very high indication that variability between providers is far in excess of variability within providers.
8. Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? NOTE: If multiple methods used, at least one must be appropriate.

   Submission document: Testing attachment, section 2a2.2

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable** (score-level testing was not performed)
9. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

   Submission document: Testing attachment, section 2a2.2

   ☐ **Yes**

   ☐ **No**

   ☒ **Not applicable** (data element testing was not performed)
10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

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

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

    ☐ **Low** (NOTE: Should rate LOW if you believe specifications are NOT precise, unambiguous, and complete or if testing methods/results are not adequate)

    ☐ **Insufficient** (NOTE: Should rate INSUFFICIENT if you believe you do not have the information you need to make a rating decision)
11. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**

    - The reliability was specified with a 10 event cut-off even as the measure was not specified as such.
    - Testing does not include sensitivity analysis on the exclusions.

- The high Adams R score alone is in no small part related to the very large sample size.
- No current Kappa stats were presented, though previous stats suggested fair to substantial reliability of the numerator and denominator based on EHRs.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

    Submission document: Testing attachment, section 2b2.

    - No concerns with measure exclusions.
    - Among the 8,458 physicians with a minimum number of 10 quality reporting events (total n= 1.66 million), the total number of reported exclusions is 9,785. The proportion of exclusions to patients overall is 0.006.

13. **Please describe any concerns you have regarding the ability to identify meaningful differences in performance.**

    Submission document: Testing attachment, section 2b4.

    - No concerns.
    - Measures of central tendency, variability and dispersion were calculated to identify meaningful differences in performance.

14. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    Submission document: Testing attachment, section 2b5.

    - N/A

15. **Please describe any concerns you have regarding missing data.**

    Submission document: Testing attachment, section 2b6.

    - No concern. PQRS dataset did not contain missing data.

16. **Risk Adjustment**

    16a. **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

    16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

    ☐ Yes ☐ No ☒ Not applicable

    16c. **Social risk adjustment:**

    16c.1 Are social risk factors included in risk model? ☐ Yes ☐ No ☒ Not applicable

    16c.2 Conceptual rationale for social risk factors included? ☐ Yes ☐ No

    16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☐ Yes ☐ No

    16d.**Risk adjustment summary:**

    16d.1 All of the risk-adjustment variables present at the start of care? ☐ Yes ☐ No

    16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☐ Yes ☐ No

    16d.3 Is the risk adjustment approach appropriately developed and assessed? ☐ Yes ☐ No

    16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☐ Yes ☐ No

    16d.5.Appropriate risk-adjustment strategy included in the measure? ☐ Yes ☐ No

    16e**. Assess the risk-adjustment approach**

VALIDITY: TESTING

17. **Validity testing level: ☒ Measure score ☐ Data element ☐ Both**
18. **Method of establishing validity of the measure score:**

    ☐ **Face validity**

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

    ☐ **N/A (score-level testing not conducted)**
19. **Assess the method(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.2

    - Conducted correlation analysis with two measures Preventive Care and Screening: Screening for High Blood Pressure and Follow-up Documented (PQRS #317) and Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan measure (PQRS #134) – hypothesis that there is a positive association between patients screened for unhealthy alcohol use and who received brief counseling and patients who were screened for high blood pressure and clinical depression and if needed a follow-up plan was documented. Validity is moreover supported by the higher correlation between depression and alcohol screening, a higher correlation supported by what is known about such disease co-occurrence rates.
20. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    - Developer found a positive correlation between the measures:

      PQRS #317

      Coefficient of correlation = 0.29

      P-value < 0.00001

      PQRS #134

      Coefficient of correlation = 0.61

      P-value < 0.00001
21. **Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships?**

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (score-level testing was not performed)
22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?** *NOTE that data element validation from the literature is acceptable.*

    Submission document*: Testing attachment, section 2b1.*

    ☐ **Yes**

    ☐ **No**

    ☒ **Not applicable** (data element testing was not performed)
23. **OVERALL RATING OF VALIDITY taking into account the results and scope of all testing and analysis of potential threats.**

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

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

☐ **Low** (NOTE: Should rate LOW if you believe that there are threats to validity and/or relevant threats to validity were not assessed OR if testing methods/results are not adequate)

☐ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the score level and the data element level is required; if not conducted, should rate as INSUFFICIENT.)

24. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**

- Score level testing was conducted. Correlation analysis demonstrated validity of the measure. However, no direct testing was conducted to verify that sensitive and specific screening was actually done, AND that positive screens were indeed followed by bonafide (i.e., high-fidelity) brief interventions.
- Previous TEP review also confirmed the measure to be strong regarding its validity.

ADDITIONAL RECOMMENDATIONS

25. **If you have listed any concerns in this form, do you believe these concerns warrant further discussion by the multi-stakeholder Standing Committee? If so, please list those concerns below.**

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

2a1. Reliability – Specifications
Comments:
**No concerns.
**Unclear what standardized screening "method" is. This could mean something that the physician makes up and has no evidence. Also unclear to what extent follow up/brief counseling is expected. Nor is the extent of the brief counseling/how to do it included and it is unlikely there will be poor fidelity and poor implementation.
**Including both abuse and dependence in the measure seems less concerning than it might have previously been given evolution of dx to DSM 5. given widespread use, no concerns that the measure can be consistently implemented.
**No concerns about reliability; it can be consistently implemented.
**No concerns, high reliability.
**Reliability appears good.

2a2. Reliability – Testing
Comments:
**No concerns.
**Reliability is likely moderate.
**No.
**No.
**This requires chart audits to determine if there is an exclusion. I think that will have a negative impact on reliability. In there discussion they identified that they did not have any patient that had an exclusion. How will this be implemented in a real world practice - they need to go into charts to determine exclusions?
**No concerns.
**No concerns, high reliability.
**No.

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

Comments:
**No concerns.
**Validity is poor for the reasons stated above regarding evidence based screening and counseling.
**No.
**No.
**No.
**No concerns.
**No.
**No.
**Missing data does not constitute a threat.
**No.
**No major concerns. various measures of "better" care seem to correlate. there does seem to be some improvement in scores over time and although average scores are low there do seem to be significant and meaningful differences between plans.
**One consideration is that often screening is considered part of the visit and cannot be billed for separately. Often items that are not billable events are not coded in the claims data.
**Issue of chart audits for exclusions is a problem in my mind.
**I do not believe it constitutes a substantial threat.
**Range suggests clinically meaningful performance variation. 2b5 not performed and 2b6 data not available.
**No big issues

2b2-3. Other Threats to Validity

2b2. Exclusions

2b3. Risk Adjustment

Comments:
** No risk adjustment is necessary.
**Disparities data not available; would be better if measure included those under 18yo.
**Exclusions (e.g., palliative care) seem reasonable. risk adjustment does not apply.
**Analyses indicate acceptable results.
**Chart audit issue.
**Risk adjustment was appropriately tested; results are acceptable.
**Measure exclusions includes patient preference but total exclusions are relatively small so do not view as threat. Risk adjustment N/A.
**It would be nice to see disparities adjusted data.

Criterion 3. Feasibility

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

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

- Data for the measure are routinely generated and used by healthcare personnel in the course of care. Some data elements are in defined fields in electronic sources
    - Additional information provided by the developer on 12.14.18 notes that while a registry is an electronic data source and registries can pull data from EHRs, the registry measure can also use claims data in the federal program.

Questions for the Committee:

- *Are there any additional considerations for implementing this measure?*

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility

Comments:

**No concerns.

**I think the lack of guidance and specificity on a particular screning tool will weaken feasibility; PC practices will have to ensure that this measure is put into EHR with reminders to rescreen every 24 months. Also, I'd like to see a continuity of care plan so that the patient will return for more than one brief intervention. I wonder what will be counted as a screening method.

**No concerns about a measure already in widespread use.

**Screening may not appear in electronic billing data.

**Is less feasible with the chart audit requirement.

**No concerns; it is feasible to collect.

**No concerns. Generated as part of care delivery and available.

**Feasibility requires EHR data field which may not be available.

Criterion 4: Usability and Use

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

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

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

4a.1. Accountability and Transparency. Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

- The measure is currently included in the Merit-based Incentive Payment System (MIPS). Prior to 2016, the measure was used in the Physician Quality Reporting System (PQRS).
    - 2018 data will be available for public reporting on Physician Compare in late 2019.

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

Feedback on the measure by those being measured or others

- The developer assembles a multistakeholder group to give input on the measure development process. Also, the developer gathers feedback from those who implement the measure via a public and member comment period or email. The developer received supportive comments for this measure, comments requesting consideration of a lower age range and comments requesting the addition of a medical reason exception for patients with limited life expectancy. Implementers also requested the developer clarify what qualifies and does not qualify as meeting the measure.
- The developer incorporated the addition of a medical reason exception for patients with limited life expectancy into the final version of the measure as a result of implementer feedback.

Additional Feedback: N/A

***Questions for the Committee**:*

- *How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Does the measure encourage advance (and not pernicious) screenings and brief interventions targeting alcohol abuse and dependence?*

Preliminary rating for Use: ☒ Pass ☐ No Pass

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

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

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

Improvement results

- This measure was first introduced in PQRS program. Performance rates from the PQRS program were relatively stable and remained low. In 2015, the measure had an average performance rate of 74.0%. The most significant variation is between the 10th and 25th percentiles, suggesting room for improvement in those low-performing programs.

4b2. Benefits vs. harms. Benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Unexpected findings (positive or negative) during implementation

N/A

Potential harms

- The developer is unaware of any unexpected findings and/or benefits from implementation of this measure.
- The developer does comment about screening burden.
- They also note reporting concerns and burden, vis-à-vis exceptions to inclusion.

Additional Feedback

- In 2015 this measure was brought before MAP to be considered in the Medicare and Medicaid EHR Incentive Program for Eligible Professionals. MAP indicated that alcohol screening and brief intervention is evidence based and encouraged further development of this measure as an eMeasure.
- In 2015 this measure was brought before MAP to be considered in the Medicare Shared Savings Program. The MAP stated that they would support this measure for MSSP only if it rolled up into a composite measure.

- In 2015 this measure was brought before MAP to be considered in the PQRS/Physician Compare/Physician Feedback/VBPM program. MAP indicated that alcohol screening and brief intervention is evidence based and encouraged further development of this eMeasure.

Questions for the Committee:

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

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

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

4a1. Use - Accountability and Transparency

Comments:

**No concerns.
**Yes, looks fine.
**Measure is in widespread use. Feedback has been incorporated, e.g., to incorporate telehealth and medical rx.
**Yes.
**Had appropriate feedback.
**Performance results available; plan for implementation is appropriate.
**Publicly reported, used in MIPS. Feedback part of the measure development process.
**In MIPS; feedback provided and considered.

4b1. Usability – Improvement

Comments:

**No concerns.
**No unintended consequences, little harm. but unclear if the benefit is long-term.
**Unintended consequences were not reported. clearly there is much room for improvement and this type of behavioral health treatment has only gotten more critical over time.
**Benefits exceed any harm.
**No issues.
**Benefits far outweigh harms. Screening is recommended by the USPSTF after a rigorous systematic review of the evidence.
**No harms.
**No evidence of harms, except opportunity costs.

Criterion 5: Related and Competing Measures

Related or competing measures

The following measures are related:

- 2599: Alcohol Screening and Follow-Up for People with Serious Mental Illness (NCQA)
- 1661: SUB-1 Alcohol Use Screening (TJC)
- 1663: SUB-2 Alcohol Use Brief Intervention Provided or Offered and SUB 2a Alcohol Use Brief Intervention (TJC)

Harmonization

The developer notes that the NCQA measure focuses on a specific sub-population (people with serious mental illness) and is intended for use at the health plan level. In The Joint Commission measures, screening and intervention are separate measures. Additionally, The Joint Commission measures are intended for use at the hospital level. The developer was contacted by these measure stewards

respectively while the measures were developed, and they are currently harmonized to the extent feasible.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

5. Related and Competing

Comments

**Harmonized to the extent possible.

**Has already been aligned with other similar/competing measures.

**There are related measures (that generally don't seem to really be competing. I'd like to hear from the delveoper about whether they think that there are potential gains still to be made in harmonizing.

**Should be harmonized with #1661 and #1663.

**No issues.

**Other measures were developed after this measure, but the NCQA measure focuses on people with serious mental illness and in the Joint Commission measures, screening and intervention are seperate measures and it is intended for hospital use.

**Related but not competing.

**Three other measures, generally harmonized.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/22/2019

- There have been no comments or support/non-support choices as of this date.

Developer Submission

Additional evaluations and submission materials attachments…

Brief Measure Information

NQF #: 2152

Corresponding Measures:

De.2. Measure Title: Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling

Co.1.1. Measure Steward: PCPI Foundation

De.3. Brief Description of Measure: Percentage of patients aged 18 years and older who were screened for unhealthy alcohol use using a systematic screening method at least once within the last 24 months AND who received brief counseling if identified as an unhealthy alcohol user.

1b.1. Developer Rationale: This measure is intended to promote unhealthy alcohol use screening and brief counseling which have been shown to be effective in reducing alcohol consumption, particularly in primary care settings.

S.4. Numerator Statement: Patients who were screened for unhealthy alcohol use using a systematic screening method at least once within the last 24 months AND who received brief counseling if identified as an unhealthy alcohol user

S.6. 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

S.8. Denominator Exclusions: Documentation of medical reason(s) for not screening for unhealthy alcohol use (eg, limited life expectancy, other medical reasons)

De.1. Measure Type: Process

S.17. Data Source: Registry Data

S.20. Level of Analysis: Clinician : Group/Practice, Clinician : Individual

IF Endorsement Maintenance – Original Endorsement Date: Mar 04, 2014 **Most Recent Endorsement Date:** Mar 04, 2014

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

2597:Substance Use Screening and Intervention Composite

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

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

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

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

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

2152_nqf_evidence_attachment__01NOV18_Final.docx

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

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

No

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 2152

Measure Title: Unhealthy Alcohol Use: Screening & Brief Counseling

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

Date of Submission: 11/1/2018

Instructions

- *Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.*
- *Complete EITHER 1a.2, 1a.3 or 1a.4 as applicable for the type of measure and evidence.*
- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Outcome: 3 Empirical data demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service. If not available, wide variation in performance can be used as evidence, assuming the data are from a robust number of providers and results are not subject to systematic bias.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: 5 a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence 4 that the measured structure leads to a desired health outcome.
- Efficiency: 6 evidence not required for the resource use component.
- For measures derived from patient reports, evidence should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.
- Process measures incorporating Appropriate Use Criteria: See NQF's guidance for evidence for measures, in general; guidance for measures specifically based on clinical practice guidelines apply as well.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.
5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.
6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

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

Outcome

☐ Outcome:

☐Patient-reported outcome (PRO):

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

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

☒ Process: The measure focuses on screening adults for unhealthy alcohol use and the provision of brief counseling for those identified as unhealthy alcohol users

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

This measure is intended to promote unhealthy alcohol use screening and brief counseling which has been shown to be effective in reducing alcohol consumption, particularly in primary care settings. Unhealthy alcohol use "contributes to hypertension, cirrhosis, gastritis, gastric ulcers, pancreatitis, breast cancer, neuropathy, cardiomyopathy, anemia, osteoporosis, cognitive impairment, depression, insomnia, anxiety, suicide, injury, and violence."(1)

Reference:

1. Jonas DE, Garbutt JC, Amick HR, et al. Behavioral Counseling After Screening for Alcohol Misuse in Primary Care: A Systematic Review and Meta-analysis for the U.S. Preventive Services Task Force. Ann Intern Med. 2012 Sep 25

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

Not applicable

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

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

Not applicable

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☒ US Preventive Services Task Force Recommendation

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

☐ Other

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	• **Title: Behavioral Counseling After Screening for Alcohol Misuse in Primary Care: A Systematic Review and Meta-analysis for the U.S. Preventive Services Task Force** • **Author: Jonas DE, Garbutt JC, Amick HR, Brown JM, Brownley KA, Council CL, Viera AJ, Wilkins TM, Schwartz CJ, Richmond EM, Yeatts J, Swinson Evans T, Wood SD, and Harris RP.** • **Date: November 6, 2012** • **Citation: Jonas DE, Garbutt JC, Amick HR, Brown JM, Brownley KA, Council CL, Viera AJ, Wilkins TM, Schwartz CJ, Richmond EM, Yeatts J, Swinson Evans T, Wood SD, and Harris RP. Behavioral counseling after screening for alcohol misuse in primary care: A systematic review and meta-analysis for the U.S. Preventive Services Task Force. Ann Intern Med. 2012;157:645-654.** • **URL: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/alcohol-misuse-screening-and-behavioral-counseling-interventions-in-primary-care**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	The USPSTF recommends that clinicians screen adults aged 18 years and older for alcohol misuse and provide persons engaged in risky or hazardous drinking with brief behavioral counseling interventions to reduce alcohol misuse.

Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	Moderate Strength of Evidence: Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of the effect and may change the estimate.
Provide all other grades and definitions from the evidence grading system	The strength of evidence was graded based on the guidance established for the AHRQ Evidence-based Practice Center Program. Developed to grade the overall strength of a body of evidence, this approach incorporates four key domains: risk of bias (includes study design and aggregate quality), consistency, directness, and precision of the evidence. We considered all evidence from intermediate outcomes to be indirect. It also considers other optional domains that may be relevant for some scenarios, such as a dose-response association, plausible confounding that would decrease the observed effect, strength of association (magnitude of effect), and publication bias. Definitions of the grades of overall strength of evidence Grade: Definition 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 the effect and may change the estimate. Low: Low confidence that the evidence reflects the true effect. Further research is likely to change our confidence in the estimate of the effect and is likely to change the estimate. Insufficient: Evidence either is unavailable or does not permit estimation of an effect.
Grade assigned to the **recommendation** with definition of the grade	B Recommendation The USPSTF recommends this service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. Source: https://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions#grade-definitions-after-july-2012
Provide all other grades and definitions from the recommendation grading system	A Recommendation: The USPSTF recommends this service. There is high certainty that the net benefit is substantial. B Recommendation: The USPSTF recommends this service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial. C Recommendation: The USPSTF recommends selectively offering or providing this service to individual patients based on professional judgement and patient preferences. There is at least moderate certainty that the net benefit is small. D Recommendation: The USPSTF recommends against this 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 that 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. Source: https://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions#grade-definitions-after-july-2012

Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	The USPSTF evidence review included 23 randomized, controlled trials included in 38 articles. The quality of the body of evidence for adults was summarized according to the grades of evidence rating as "moderate strength of evidence" for each of 3 intermediate outcomes reported [consumption (mean drinks/week), heavy drinking episodes, and achievement of recommended drinking limits]. For specific patient populations, the quality of the body of evidence varied depending on the intermediate outcome studied. Details are as follows: Adults: Consumption: Reduction of 3.6 (2.4 to 4.8) from baseline ~23 [Moderate Strength Of Evidence (SOE)] Heavy Drinking Episodes: 12% fewer subjects reported heavy drinking episodes (7%, 16%) from ~52% at baseline [Moderate SOE] Recommended Drinking Limits: 11% more subjects achieved (8%, 13%) [Moderate SOE] Older adults: Consumption: Reduction of 1.7 (0.6 to 2.8) from baseline ~16 [Moderate SOE] Heavy Drinking Episodes: [Insufficient SOE] Recommended Drinking Limits: 9% more subjects achieved (2%, 16%) [Low SOE] Young adults or college students Consumption: Reduction of 1.7 (0.7 to 2.6) from baseline ~15 [Moderate SOE] Heavy Drinking Episodes: 0.9 fewer heavy drinking days (0.3, 1.5) from ~6.2 days per month at baseline [Moderate SOE] Recommended Drinking Limits: [Insufficient SOE] Pregnant women Consumption: Data from 1 study found no difference [Low SOE] Heavy Drinking Episodes: [Insufficient SOE] Recommended Drinking Limits: [Insufficient SOE] Of note, none of the studies were designed to achieve abstinence, and the report indicated it should probably not be a goal of behavioral interventions for most people. For most [long term] health outcomes, available evidence either demonstrated no difference between interventions and controls (e.g., mortality: low SOE) or was insufficient to draw conclusions (e.g., accidents, injuries, alcohol-related liver problems: insufficient SOE). Some evidence suggests that interventions improve some utilization outcomes for adults (e.g., hospital days and costs: low SOE). [The recent] meta-analyses did not find a reduction in all-cause mortality for adults (four studies; rate ratio 0.64, 95% confidence interval [CI], 0.24 to 1.7) or for all age groups combined (adults, older adults, and young adults/college students) (six studies; rate ratio 0.52, 95% CI, 0.22 to 1.2).

Estimates of benefit and consistency across studies	Although the results by population group are summarized in the section above, additional details addressing the consistency of results across studies are provided below: --Consumption: Behavioral interventions resulted in a greater reduction in quantity of alcohol consumed than controls at 12 months (weighted mean difference [WMD], -3.6 drinks per week, 95% CI, -4.8 to -2.4, moderate SOE). Subgroup analyses for men and women found similar benefits. When stratifying by intensity of the intervention, we found no statistically significant difference between very brief interventions and controls (just one study contributed), but found greater reduction for brief, brief multi-contact, and extended multi-contact interventions than for controls. We found similar results for studies conducted in the United States compared with those conducted in other countries, a trend toward a greater reduction in consumption for interventions delivered primarily by primary care providers (WMD, -4.0 drinks per week, 95% CI, -5.4 to -2.6) than for those delivered primarily by research personnel (WMD, -3.0, 95% CI, -5.0 to -1.0), and that studies enrolling 10 percent or more subjects with alcohol dependence found behavioral interventions to be ineffective or less effective than other studies. -- Heavy drinking episodes: Behavioral interventions resulted in 12 percent more subjects reporting no heavy drinking episodes by 12 months compared with controls (risk difference 0.12, 95% CI, 0.07 to 0.16, moderate SOE). Subgroup analyses for men and women found similar results. When stratifying by intensity of the intervention, brief multi-contact and extended multi-contact interventions were efficacious at 12 months (with 11 percent and 19 percent absolute difference compared with controls, respectively), but brief interventions did not reach statistical significance compared with controls. -- Recommended drinking limits achieved: 11 percent more subjects receiving interventions achieved recommended drinking limits by 12 months compared with controls (risk difference 0.11, 95% CI, 0.08 to 0.13, moderate SOE). Subgroup analyses for men and women found similar magnitude of benefit. All of the intervention intensities studied were efficacious. The absolute difference in percentage of subjects achieving recommended drinking limits was numerically greatest for the brief multi-contact interventions (15% compared with 8%for very brief and brief interventions at 12 months), but the confidence intervals overlap.
What harms were identified?	The study authors found no evidence of direct harms, aside from opportunity costs associated with interventions, which ranged from 5 minutes to 2 hours dispersed over several in-person or telephone visits [moderate SOE]. The authors searched for evidence of potential adverse effects, such as illegal substance use, increased smoking, anxiety, stigma, labeling, discrimination, or interference with the physician–patient relationship. They found no evidence for most of these potential harms and very limited evidence reporting no difference between groups for smoking rates and anxiety [low SOE]. Other than the results for opportunity costs, the results are limited by the few trials that reported any information; 5 of 23 reported smoking, and 2 reported anxiety.

Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	We are aware that the current guideline recommendation is under review by the USPSTF. The public comment version of the draft recommendation statement is posted at: https://www.uspreventiveservicestaskforce.org/Page/Document/final-research-plan/unhealthy-alcohol-use-in-adolescents-and-adults-including-pregnant-women-screening-and-behavioral-counseling-interventions The public comment period ended on July 2, 2018. A date for when to expect the final updated guideline recommendation to be published is not yet known. Upon review of the draft recommendation statement, there are no changes that would affect the measure.

1a.4 OTHER SOURCE OF EVIDENCE

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

Not applicable.

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

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

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

1b. Performance Gap

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

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

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

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

This measure is intended to promote unhealthy alcohol use screening and brief counseling which have been shown to be effective in reducing alcohol consumption, particularly in primary care settings.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis*. (This is required for maintenance of endorsement. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

An abbreviated version of this measure (which focuses only on the screening for unhealthy alcohol use component of the measure and not the brief counseling component) was included in CMS' Physician Quality Reporting System (PQRS) program from 2009-2016. Average performance rates from 2012 through 2015, reflecting the most recent data that have been made publicly available, are as follows:

PQRS #173 Preventive Care and Screening: Unhealthy Alcohol Use – Screening

2012: 74.5%

2013: 75.5%

2014: 66.2%

2015: 74.0%

Additional data analysis provided for 2015 PQRS data are as follows:

Performance 10th Percentile: 19.80

Performance 25th Percentile: 56.60

Performance 50th Percentile: 84.62

Performance 75th Percentile: 100.00

Performance 100th Percentile: 100.00

Performance Interquartile Range: 43.41 (2)

The current version of the measure is included in the Merit-based Incentive Payment System (MIPS), however data are not yet available.

1. Centers for Medicare & Medicaid Services. Physician Quality Reporting System 2015 Reporting Experience Including Trends (2007-2015). 2017. Available at:https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/AnalysisAndPayment.html

2. Additional 2015 PQRS data provided as requested from CMS.

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, including patient and provider surveys, have documented low rates of alcohol misuse screening and counseling in primary care settings.

According to a study analyzing the quality of health care in the United States, on average, 45% of patients (n=6,676) were screened for problem drinking.(1)

In the national Healthcare for Communities Survey, only 8.7% of problem drinkers reported having been asked and counseled about their alcohol use in the last 12 months.(2)

A nationally representative sample of 648 primary care physicians were surveyed to determine how such physicians identify--or fail to identify--substance abuse in their patients, what efforts they make to help these patients with such morbidity and what are the barriers to effective diagnosis and treatment. Of physicians who conducted annual health histories, less than half ask about the quantity and frequency of alcohol use (45.3 percent). Only 31.8 percent say they ever administer standard alcohol or drug use screening instruments to patients. (3)

A national systematic sample of 2,000 physicians practicing general internal medicine, family medicine, obstetrics-gynecology, and psychiatry were surveyed to determine the frequency of screening and intervention for alcohol problems. Of the 853 respondent physicians, 88% usually or always ask new outpatients about alcohol use. When evaluating patients who drink, 47% regularly inquire about maximum amounts on an occasion, and 13% use formal alcohol screening tools. Only 82% routinely offer intervention to diagnosed problem drinkers. (4)

In 2014, the CDC analyzed data from 17 states and the District of Columbia via the Behavioral Risk Factor Surveillance System to estimate the prevalence of adults who reported receiving elements of alcohol screening and brief intervention. While 77.7% of adults reported being asked about alcohol use by a health professional, only 32.9% were asked about binge-level alcohol consumption and among binge drinkers only 37.2% reported being counseled on the harms of binge drinking. Only 18.1% reported being advised to cut down on alcohol consumption or to quit drinking. (5)

A multi-site, cross-sectional survey of primary care residents from six primary care residency programs administered from March 2010 through December 2012 found that a minority of the residents appropriately screen or provide intervention for at risk alcohol users. While 60% (125/208) stated they screen patients at an initial visit, only 17% (35/208) screened patients at subsequent visits. 54% (108/202) reported they did not feel

they had adequate training to provide brief intervention to patients found to be at-risk alcohol users and 21% (43/208) felt they could really help at-risk drinkers. (6)

A study evaluating self-reported prevalence of alcohol screening using information drawn from the ConsumerStyles survey (a random internet panel) found that only 24.7% (n=2,592) of adults reported being asked about their alcohol use While prevalence among men and women were about the same, there was lower prevalence of screening among Black non-Hispanics than white non-Hispanics (16.2% vs. 26.9%) and college graduates reported a higher prevalence of screening than those with a high school degree or less (38.1% vs. 20.8%). (7)

A cross-sectional analysis using 2016 DocStyles data that evaluated with use of different screening tools used to screen for alcohol misuse by 1,506 primary care providers found that while most providers screen for alcohol misuse (96%) only 38% reported using a USPSTF recommended screening tool. (8)

1. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med. 2003;348:2635-2645.

2. D´Amico EJ, Paddock SM, Burnam A, Kung FY. Identification of and guidance for problem drinking by general medical providers: results from a national survey. Med Care. 2005 Mar;43(3):229-36.

3. Missed Opportunity: National Survey of Primary Care Physicians and Patients on Substance Abuse. New York: The National Center on Addiction and Substance Abuse at Columbia University; 2000.

4. Friedmann PD, McCullough D, Chin MH, Saitz R. Screening and intervention for alcohol problems. A national survey of primary care physicians and psychiatrists. J Gen Intern Med. 2000 Feb;15(2):84-91.

5. McKnight-Eily LR, Okoro CA, Mejia R, Denny CH, Higgins-Biddle J, Hungerford D, et al. Screening for excessive alcohol use and brief counseling of adults—17 states and the District of Columbia, 2014. MMWR Morb Mortal Wkly Rep 2017;66:313-319.

6. Barnes Le K, Johnson A, Seale P, Woodall H, Clark DC, Parish DC, et al. Primary care residents lack comfort and experience with alcohol screening and brief intervention: A multi-site survey. J Gen Intern Med. 2015. 30(6):790-6.

7. Denny CH, Hungerford DW, McKnight-Elly LR, Green PP, Dang Ep, Cannon MJ, et al. Self-reported prevalence of alcohol screening among U. S. adults. Am J Prev Med. 2016. March;50(3):380-383.

8. Tan CH, Hungerford DW, Denny C, McKnight-Eily LR. Screening for alcohol misuse: Practices among U.S. primary care providers, DocStyles 2016. Am J Prev Med. 2018;54(2):173-180.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for maintenance of endorsement. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

While this measure is included in a federal reporting program, disparities data have not yet been made available to 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

Historically, literature has shown variations across race and "ethnicities in drinking, alcohol use disorders, alcohol problems, and treatment use. Higher rates of high-risk drinking among ethnic minorities are reported for Native Americans and Hispanics, although within ethnic group differences (e.g., gender, age group, and other subpopulations) also are evident for ethnicities. Whites and Native Americans have a greater risk for alcohol use disorders relative to other ethnic groups. However, once alcohol dependence occurs, Blacks and

Hispanics experience higher rates than Whites of recurrent or persistent dependence. Furthermore, the consequences of drinking appear to be more profound for Native Americans, Hispanics, and Blacks."(1)

More recent literature shows that there are differences in patient populations that are receive screening for unhealthy alcohol use.

A study evaluating self-reported prevalence of alcohol screening using information drawn from the ConsumerStyles survey (a random internet panel) found that only 24.7% (n=2,592) of adults reported being asked about their alcohol use While prevalence among men and women were about the same, there was lower prevalence of screening among Black non-Hispanics than white non-Hispanics (16.2% vs. 26.9%) and college graduates reported a higher prevalence of screening than those with a high school degree or less (38.1% vs. 20.8%). (2)

In 2014, the CDC analyzed data from 17 states and the District of Columbia via the Behavioral Risk Factor Surveillance System to estimate the prevalence of adults who reported receiving elements of alcohol screening and brief intervention. The prevalence of being asked about binge drinking was higher among males (35%) and in people with less than a high school diploma (40.1%). Additionally, non-Hispanic whites and Asian/Pacific Islanders were asked about binge drinking less frequently than non-Hispanic blacks and American Indian/Alaskan Natives. (3)

1. Chartier K, Caetano R. Ethnicity and Health Disparities in Alcohol Research. Alcohol Res Health. 2010;33(1-2):152-160.

2. Denny CH, Hungerford DW, McKnight-Elly LR, Green PP, Dang Ep, Cannon MJ, et al. Self-reported prevalence of alcohol screening among U. S. adults. Am J Prev Med. 2016. March;50(3):380-383.

3. McKnight-Eily LR, Okoro CA, Mejia R, Denny CH, Higgins-Biddle J, Hungerford D, et al. Screening for excessive alcohol use and brief counseling of adults—17 states and the District of Columbia, 2014. MMWR Morb Mortal Wkly Rep 2017;66:313-319.

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 : Alcohol, Substance Use/Abuse

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

Primary Prevention, Screening

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

Elderly

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

The measure specifications are included as an attachment with this submission. Additional measure details may be found at the PCPI website: http://www.thepcpi.org/?page=PCPIMeasures

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

This is not an eMeasure **Attachment:**

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

No data dictionary **Attachment:**

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

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

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

Not an instrument-based measure

S.3.1. <u>For maintenance of endorsement:</u> 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. <u>For maintenance of endorsement,</u> 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. However, this annual review has not resulted in any changes for this measure.

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

<u>IF an OUTCOME MEASURE</u>, 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 unhealthy alcohol use using a systematic screening method at least once within the last 24 months AND who received brief counseling if identified as an unhealthy alcohol 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)*

<u>IF an OUTCOME MEASURE,</u> 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:

Systematic screening method - For purposes of this measure, one of the following systematic methods to assess unhealthy alcohol use must be utilized. Systematic screening methods and thresholds for defining unhealthy alcohol use include:

- AUDIT Screening Instrument (score >= 8)
- AUDIT-C Screening Instrument (score >= 4 for men; score >= 3 for women)
- Single Question Screening - How many times in the past year have you had 5 (for men) or 4 (for women and all adults older than 65 years) or more drinks in a day? (response >= 2)

Brief counseling - Brief counseling for unhealthy alcohol use refers to one or more counseling sessions, a minimum of 5-15 minutes, which may include: feedback on alcohol use and harms; identification of high risk situations for drinking and coping strategies; increased motivation and the development of a personal plan to reduce drinking.

NUMERATOR NOTE: In the event that a patient is screened for unhealthy alcohol use and identified as a user but did not receive brief alcohol cessation counseling submit G9624.

For Registry:

Report Quality Data Code:

G9621 - Patient identified as an unhealthy alcohol user when screened for unhealthy alcohol use using a systematic screening method and received brief counseling

OR

G9622 - Patient not identified as an unhealthy alcohol user when screened for unhealthy alcohol use using a systematic screening method

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 Registry:

Patients aged >= 18 years

AND

At least two patient encounters during the performance period (CPT or HCPCS): 90791, 90792, 90832, 90834, 90837, 90845, 96150, 96151, 96152, 97165, 97166, 97167, 97168, 97802, 97803, 97804, 99201, 99202, 99203, 99204, 99205, 99212, 99213, 99214, 99215, G0270, G0271

WITHOUT

Telehealth Modifier: GQ, GT, 95, POS 2

OR

At Least One Preventive Visit during the performance period (CPT or HCPCS): 99385, 99386, 99387, 99395, 99396, 99397, 99401, 99402, 99403, 99404, 99411, 99412, 99429, G0438, G0439

WITHOUT

Telehealth Modifier: GQ, GT, 95, POS 02

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

Documentation of medical reason(s) for not screening for unhealthy alcohol use (eg, limited life expectancy, other medical reasons)

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: Denominator Exception(s) are determined on the date of the most recent denominator eligible encounter.

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: Unhealthy Alcohol Use: Screening & Brief Counseling, exceptions may include medical reason(s) (eg, limited life expectancy, other medical reasons). 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 Registry:

Report Quality Data Code:

G9623 - Documentation of medical reason(s) for not screening for unhealthy alcohol use (e.g., limited life expectancy, other medical reasons)

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

Consistent with CMS' Measures Management System Blueprint and recent national recommendations put forth by the IOM and NQF, the PCPI encourages the collection of race and ethnicity data as well as the results of this measure to be stratified by race, ethnicity, administrative sex, and payer.

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
1. 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: medical reason(s) (eg, limited life expectancy, other medical reasons)]. 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 an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Not applicable. The measure does not require sampling.

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

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

Not applicable. This measure does not use a survey or an instrument.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Registry 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 are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

Not applicable.

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

No data collection instrument provided

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

Clinician : Group/Practice, Clinician : Individual

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

Home Care, Outpatient Services

If other:

S.22. COMPOSITE Performance Measure - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

Not applicable.

2. Validity – See attached Measure Testing Submission Form

2152_nqf_testing_attachment_7.1.docx

2.1 For maintenance of endorsement

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

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 For maintenance of endorsement

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

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 2152
Measure Title: Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling
Date of Submission: 8/1/2018

Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. *If there is more than one set of data specifications or more than one level of analysis, contact NQF staff* about how to present all the testing information in one form.
- For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.
- For outcome and resource use measures, section 2b3 also must be completed.
- If specified for multiple data sources/sets of specificaitons (e.g., claims and EHRs), section 2b5 also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this

form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.

- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). *Contact NQF staff if more pages are needed.*
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.

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

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

2b1. Validity testing 11 demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument-based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

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

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). 13

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

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

OR

- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

Notes

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

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid

indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.

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

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

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

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

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

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

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

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

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

The data source is Electronic Health Records and Registry data

The data source is 2016 Registry 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?

The measurement period (data collected from patients seen) was 8/1/2011 through 7/31/2012.

The data are for the time period January 2016 through December 2016 and cover the entire United States.

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

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

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

This study captured performance on 97 events, the data were not captured at the physician level, restricting reporting of variation in performance to the organization level only.

We received data from 9,511 physicians reporting on this measure through the registry option for CMS's PQRS in 2016. Of those, 8,458 physicians had all the required data elements and met the minimum number of quality reporting events (10) for our analysis for a total of 1,660,749 quality events. For this measure, 89 percent of physicians are included in the analysis, and the average number of quality reporting events are 196 for the remaining 1,660,749 events. The range of quality reporting events for 8,458 physicians included is from 10 to 5,579. The average number of quality reporting events for the remaining 11 percent of physicians that aren't included is 4.

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

There were 1,660,749 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 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 samples were used for reliability testing and exceptions analysis.

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

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

2a2. RELIABILITY TESTING

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

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

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

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

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

Data analysis included: Percent agreement; and Kappa statistic to adjust for chance agreement

Reliability of the computed measure score was measured as the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in physician performance and the noise is the total variability in measured 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 evaluated by averaging over physician specific reliabilities for all providers that meet the minimum number of quality reporting events for the measure. Each provider must have at least 10 eligible reporting events to be included in this calculation.

A reliability equal to zero implies that all the variability in a measure is attributable to measurement error. A reliability equal to one implies that all the variability is attributable to real differences in physician performance. A reliability of 0.70 – 0.80 is generally considered the acceptable threshold for reliability, 0.80 – 0.90 is considered high reliability, and 0.90 – 1.0 is considered very high. [1]

1. Adams JL, Mehrotra A, McGlynn EA, Estimating Reliability and Misclassification in Physician Profiling, Santa Monica, CA: RAND Corporation, 2010. www.rand.org/pubs/technical_reports/TR863. (Accessed on February 24, 2012.)

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*)

Kappa statistics were calculated at the measure level for the denominator and numerator categories as a method of analyzing the reliability of measure implementation at the testing site. For analysis of the Kappa Statistics, the AMA-PCPI used Landis, J.R.; & Koch, G.G. (1977). "The measurement of observer agreement for categorical data". Biometrics 33 (1): 159–174.

0–.20 = slight agreement

.21–.40 = fair agreement

.41–.60 = moderate agreement

.61–.80 = substantial agreement

.81–1 = almost perfect agreement

N, % agreement, kappa statistic ,(95% confidence interval)

Denominator Reliability: 120, 85.0%, 0.31 (0.10 – 0.52)

Of the 120 observations that were initially selected, 97 observations met the criteria for inclusion in the numerator analysis.

N, % agreement, kappa statistic ,(95% confidence interval)

Numerator Reliability: 97, 91%, 0.82 (0.70 – 0.93)

The reliability above the minimum level of quality reporting events was 0.99.

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 kappa statistic value of 0.31 demonstrates fair agreement. This is due to the high observed agreement rate and the concentration of observations in the YES, YES cell (81% of all observations (97/120)). This is an example of the limitation of the Kappa statistic. While agreement can be high, if one classification category dominates, kappa can be significantly reduced. (Warrens MJ, A Formal Proof of a Paradox Associated with Cohen's Kappa. Journal of Classification. 27:322-332, 2010; Feinstein AR, Cicchetti DV. High Agreement but Low Kappa: I. The Problems of Two Paradoxes. Journal of Clinical Epidemiology. 43:543–549, 1990)

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

2b1. VALIDITY TESTING

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

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Face Validity:

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

Preventive Care and Screening: Screening for High Blood Pressure and Follow-up Documented (PQRS #317) as well as the Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan measure

(PQRS #134) were chosen as suitable candidates for correlation analysis due to the similarities in patient population and domain. We hypothesize that there exists a positive association between patients aged 18 years and older who were screened for unhealthy alcohol use using a systemic screening method and who received brief counseling if identified as an unhealthy alcohol user and those who were screened for high blood pressure and a recommended follow-up plan is documented based on the current blood pressure reading as indicated. Additionally we hypothesize that there exists a positive association between patients aged 18 years and older who were screened for unhealthy alcohol use using a systemic screening method and who received brief counseling if identified as an unhealthy alcohol user and those who were screened for clinical depression using an age appropriate standardized tool and, if positive, a follow-up plan is documented on the date of the positive screen.

Providers included in the analysis met the minimum number of quality reporting events (10) and were cleaned in the same process as the PQRS dataset.

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

We use the following guidance to describe correlation[1]:

Correlation	Interpretation
> 0.40	Strong
0.20 - 0.40	Moderate
< 0.20	Weak

1. Shortell T. An Introduction to Data Analysis & Presentation. Sociology 712. http://www.shortell.org/book/chap18.html. Accessed July 13, 2018.

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

Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling was positively correlated with the Preventive Care and Screening: Screening for High Blood Pressure and Follow-up Documented measure (PQRS #317) as well as the Preventive Care and Screening: Screening for Clinical Depression and Follow-Up Plan measure (PQRS #134):

PQRS #317

Coefficient of correlation = 0.29

P-value < 0.00001

PQRS #134

Coefficient of correlation = 0.61

P-value < 0.00001

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

The results of the expert panel rating of the validity statement were as follows: N = 19; Mean rating = 4.32 and 84.2% 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 - 2

3 - 1 (Neither Agree nor Disagree)

4 - 3

5 - 13 (Strongly Agree)

Preventive Care and Screening: Unhealthy Alcohol Use: Screening & Brief Counseling has a moderate positive correlation and a strong positive correlation with other evidence-based process of care measures focused on preventive care services. The correlations demonstrate the criterion validity of the measure.

_____

2b2. EXCLUSIONS ANALYSIS

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

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

Exceptions included a medical reason. Exceptions were analyzed for frequency and variability across providers.

Exceptions include:

- Documentation of medical reason(s) for not screening for unhealthy alcohol use (e.g., limited life expectancy, other medical reasons)

Exceptions were analyzed for frequency across providers.

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

Although specifications allowed for documented medical exceptions for this measure, there were no documented exceptions in this project. All sampled patients were able to be assessed.

Amongst the 8,458 physicians with the minimum (10) number of quality reporting events, there were a total of 9,785 exceptions reported. 5% of physicians reported an exception and the average number of exceptions reported by those physicians is 22.6. The proportion of exceptions to patients overall is 0.006.

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

Exceptions are necessary to account for those situations when there is documentation of a medical reason for not screening for unhealthy alcohol use. 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 several 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 medical reason for the decision to not screen for unhealthy alcohol 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.

Although the rates of exception reporting were low, exceptions are necessary to account for those situations when there is documentation of a medical reason for not screening for unhealthy alcohol use. 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.

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

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

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with _risk factors**

☐ **Stratification by _risk categories**

☐ **Other,**

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

Not applicable

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

Not applicable

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

Not applicable

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

☐ **Published literature**

☐ **Internal data analysis**

☐ **Other (please describe)**

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

Not applicable

2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of*

unique variation in the outcome, assessment of between-unit effects and within-unit effects.) **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

Not applicable

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

Not applicable

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

If stratified, skip to 2b3.9

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

Not applicable

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

Not applicable

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

Not applicable

2b3.9. Results of Risk Stratification Analysis:

Not applicable

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

Not applicable

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

Not applicable

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

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

Data analysis performed on the measure included:

Average measure performance rate overall and by site, performance rate range by site and overall standard deviation for the measure.

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

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

Average Measure performance rate without exceptions: N= 97 Mean = 55.6% Standard Deviation= 0.4993

The performance rate by site is as follows, where n is the number of performance events by site:

1 1.000 n=40

2 1.000 n=12

3 0.044 n=45

The performance rate range is 0.956.

Based on the sample of 8,458 included physicians, the mean performance rate is 0.67 the median performance rate is 0.81 and the mode is 1.0. The standard deviation is 0.35. The range of the performance rate is 1.00, with a minimum rate of 0.00 and a maximum rate of 1.00. The interquartile range is 0.61 (0.97–0.36).

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

Although this study captured performance on 97 events, the data

were not captured at the physician level, restricting reporting of variation in performance to the organization level only. Additionally, we are unable to present a meaningful calculation of variation in performance across organizations due to the small sample size of sites (n=3) in this study.

The range of performance from 0.00 to 1.00 suggests there's clinically meaningful variation across physicians' performance.

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

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

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

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

This test was not performed for this measure.

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

This test was not performed for this measure.

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

This test was not performed for this measure.

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or

differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

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

This test was not performed for this measure. There was no missing data.

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

The PQRS dataset provided to us by CMS did not contain missing data so this test was not performed. Nevertheless, missing data may have been rejected when submitted to CMS in which case those values would not be counted towards measure performance. There is no indication that this missing data was systematic, thus their omission would lead to unbiased performance results.

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), 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).

3b.3. If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL. Please also complete and attach the NQF Feasibility Score Card.

Attachment:

3c. Data Collection Strategy

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

3c.1. Required for maintenance of endorsement. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

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

We have not identified any areas of concern or made any modifications as a result of feasibility 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)*.

Not applicable.

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 Merit-based Incentive Payment System https://qpp.cms.gov/mips/quality-measures Payment Program https://qpp.cms.gov/mips/quality-measures Merit-based Incentive Payment System

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

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

Merit-based Incentive Payment System (MIPS)-Sponsored by the Centers for Medicare and Medicaid Services (CMS)
Prior to 2016, this measure was used for Eligible Providers (EPs) in the Physician Quality Reporting System (PQRS). As of 2017, PQRS has been replaced by the Merit-based Incentive Payment System (MIPS). MIPS is a national performance-based payment program that uses performance scores across several categories to determine payment rates for EPs. MIPS takes a comprehensive approach to payment by basing consideration of quality on a set of evidence-based measures that were primarily developed by clinicians, thus encouraging improvement in clinical practice and supporting advances in technology that allow for easy exchange of information.
According to the CY 2018 Quality Payment Program final rule, CMS intends to "make all measures under MIPS quality performance category available for public reporting on Physician Compare in the transition year of the Quality Payment Program, as technically feasible." These measures include those reported via all available submission methods for MIPS-eligible clinicians and groups. Because this measure has been in use for at least one year and meets the minimum sample size requirement for reliability, this measure meets criteria for public reporting. 2018 data will be available for public reporting on Physician Compare in late 2019.

4a1.2. 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.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Not applicable.

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

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

The PCPI measure development process is a rigorous, evidence-based process that has been refined and standardized over the past fifteen years, since the PCPI's inception. Throughout its tenure, several key principles have guided the development of performance measures by the PCPI, including the following which

underscore the role those being measured have played in the development process and later through implementation feedback:

Collaborative Approach to Measure Development

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

Conduct Public Comment Period

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

Feedback Mechanism

The PCPI has a dedicated process set up to receive comments and questions from implementers. As comments and questions are received, they are shared with appropriate staff for follow up. If comments or questions require expert input, these are shared with the PCPI's expert 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.

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

See description in 4a2.1.1 above.

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

Describe how feedback was obtained.

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

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

The 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.

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

See summary in 4a2.2 above.

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

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.

Improvement

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

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

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

While average performance rates from the PQRS program seemed relatively stable, they remain low. It is important to note that PQRS, now the Merit-based Incentive Payment System (MIPS), has been and remains a voluntary reporting program. In the early years of the PQRS program, participants received an incentive for satisfactorily reporting. As a result, performance rates may not be nationally representative. Beginning in 2015, the program imposed payment penalties for non-participants based on 2013 performance.

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.

4b2. Unintended Consequences

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

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

We are not aware of any positive or negative unexpected findings for this measure.

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

We are not aware of any unexpected benefits from implementation of this measure.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

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

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

#2599: Alcohol Screening and Follow-Up for People with Serious Mental Illness (NCQA)

#1661: SUB-1 Alcohol Use Screening (TJC)

#1663: SUB-2 Alcohol Use Brief Intervention Provided or Offered and SUB 2a Alcohol Use Brief Intervention (TJC)

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 related measures listed in 5.1b were developed after our measure. The NCQA measure focuses on a specific sub-population (people with serious mental illness) and is intended for use at the health plan level. In the TJC measures, screening and intervention are separate measures. Additionally, the TJC measures are intended for use at the hospital level. PCPI was contacted by these measure stewards respectively while the measures were developed, and they are currently harmonized to the extent feasible.

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 NQF-endorsed measure.

Appendix

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

No appendix Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): PCPI Foundation

Co.2 Point of Contact: Samantha, Tierney, Samantha.Tierney@ama-assn.org, 312-224-6071-

Co.3 Measure Developer if different from Measure Steward: PCPI Foundation

Co.4 Point of Contact: Samantha, Tierney, samantha.tierney@ama-assn.org, 312-224-6071-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

PCPI measures are developed through cross-specialty, multi-disciplinary technical expert panels (TEPs). Representatives of all relevant disciplines of medicine and other health care professionals are invited to participate. In addition, the PCPI strives to include on its TEPs individuals representing the perspectives of patients, consumers, private health plans, and employers. Measure methodologists, and coding and informatics experts also are considered important members of the TEP. All TEP members participate as equal contributors to the measure development process. This broad-based approach to measure development ensures buy-in on the measures from all stakeholders and minimizes bias toward any individual specialty or stakeholder group. TEPs were convened in 2001 and 2008 to develop, refine and maintain a set of measures addressing preventive care and screening including measure #2152. More recently, in 2016, the PCPI reconvened the Preventive Care TEP which included the following individuals:

Deanna Willis MD, MBA (co-chair)

John Wong MD (co-chair)

Susan Blank MD

Joel Brill MD

Peter Briss MD

Sandra Dunbar PhD, RN

Yngve Falck-Ytter MD

Susan Friedman MD, MPH

Marc Ghany MD, MHSc

Ellen Giarelli EdD, RN, MS, CRNP

Ashley Halle OTD, OTR/L

Selena Hariharan MD

Lori Karan MD

Martin Mahoney MD, PhD

Stephen Persell MD, MPH

Brian Svazas MD, MPH

Tim Petito OD

Barbara Resnick PhD, RN, CRNP

Paola Ricci MD

Andrew Saxon MD

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: 03, 2018

Ad.4 What is your frequency for review/update of this measure? Coding/Specifications updates occur annually. See additional information below.

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

Ad.6 Copyright statement: Physician Performance Measures (Measures) and related data specifications, developed by the Physician Consortium for Performance Improvement® (the Consortium), are intended to facilitate quality improvement activities by physicians.

These Measures are intended to assist physicians in enhancing quality of care. Measures are designed for use by any physician who manages the care of a patient for a specific condition or for prevention. These performance Measures are not clinical guidelines and do not establish a standard of medical care. The Consortium has not tested its Measures for all potential applications. The Consortium encourages the testing and evaluation of its Measures.

Measures are subject to review and may be revised or rescinded at any time by the Consortium. The Measures may not be altered without the prior written approval of the Consortium. Measures developed by the Consortium, while copyrighted, can be reproduced and distributed, without modification, for noncommercial purposes, e.g., use by health care providers in connection with their practices. Commercial use is defined as the sale, license, or distribution of the Measures for commercial gain, or incorporation of the Measures into a product or service that is sold, licensed or distributed for commercial gain. Commercial uses of the Measures require a license agreement between the user and American Medical Association, on behalf of the Consortium. Neither the Consortium nor its members shall be responsible for any use of these Measures.

THE MEASURES 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 Consortium and its members disclaim all liability for use or accuracy of any Current Procedural Terminology (CPT®) or other coding contained in the specifications.

THE SPECIFICATIONS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND.

Ad.7 Disclaimers: See copyright statement above.

Ad.8 Additional Information/Comments: The PCPI has a formal measurement review process that stipulates regular (usually on a three-year cycle, when feasible) review of the measures. The process can also be

activated if there is a major change in scientific evidence, results from testing or other issues are noted that materially affect the integrity of the measure.