NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 3489

Measure Title: Follow-Up After Emergency Department Visit for Mental Illness

Measure Steward: National Committee for Quality Assurance

Brief Description of Measure: The percentage of emergency department (ED) visits for members 6 years of age and older with a principal diagnosis of mental illness or intentional self-harm, who had a follow-up visit for mental illness. Two rates are reported:

- The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days).

- The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days).

Developer Rationale: This measure targets individuals with mental health who are discharged to the community from the emergency department. These individuals may be particularly vulnerable to losing contact with the health care system. High use of the emergency department may signal a lack of access to ongoing care or a gap in fulfilling urgent care needs. Therefore, this point of transition presents an opportunity to ensure that the patient is connected to care and receives follow-up. Health plans have access to information and care management processes to ensure that follow-up care occurs. Therefore, health plans can help connect patients into outpatient care after emergency department use.

Estimates suggest that about half of psychiatric patients discharged from the emergency room transitioned successfully to outpatient care (Bruffaerts, 2005). Low-intensity interventions that can be applied widely are typically implemented at periods of high risk for treatment dropout, such as following an emergency room discharge or the time of entry into outpatient treatment (Kreyenbuhl, 2009).

Individuals discharged from the emergency department face two main risks: (1) disengagement from treatment and (2) readmission to the emergency department. Treatment disengagement is a problem because individuals with the most serious mental health problems or alcohol or drug use disorders may require ongoing support and counseling to live independently in the community. A retrospective chart study of 390 patients assessed the likelihood of patients presenting to the ED after engagement with a transitional psychiatric clinic post ED visit. Patients who had appointments in the clinic within 3 days of being discharged from the ED were more likely to be engaged in their post ED care and stayed in the community longer before returning to the ED (McCullumsith, 2015). Appropriate follow up after ED visit for mental health is needed to improve patient outcomes and treatment adherence.

Bruffaerts R, Sabbe M. Demyffenaere K. (2005) Predicting Community Tenure in Patients with Recurrent Utilization of a Psychiatric Emergency Service. Gen Hosp Psychiatry. 27:269-74.

Kreyenbuhl, J, Nossel, I, Dixon, L. (2009) Disengagement from Mental Health Treatment among Individuals with Schizophrenia and Strategies for Facilitating Connections to Care: A Review of the literature. Schizophrenia Bulletin. 35:696-703.

McCullumsmith, C., Clark, B., Blair, C., Cropsey, K., & Shelton, R. (2015). Rapid follow-up for patients after psychiatric crisis. Community mental health journal, 51(2), 139-144.

Numerator Statement: The numerator consists of two rates:

- 30-day follow-up: The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days).

- 7-day follow-up: The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days).

Denominator Statement: Emergency department (ED) visits for members 6 years of age and older with a principal diagnosis of mental illness or intentional self-harm on or between January 1 and December 1 of the measurement year.

Denominator Exclusions: Patients in hospice.

Measure Type: Process

Data Source: Claims

Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: As 2506 (with SUD) 3/6/15 **Most Recent Endorsement Date:** As 2506 (with SUD) 11/29/18

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

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

IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? Not applicable.

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.

Evidence provided by the Developer is based review of several guidelines as follows: 1. Those pertaining to psychosis and schizophrenia treatment from NICE (2014), 2. Schizophrenia from APA (2004), 3. Bipolar Disorder from APA (2002), 4. Major Depression from APA (2010). Here is a careful deconstruction of the evidence table presented under section 1a.3 of the evidence form:

1. The 1st row of the table simply lists the 4 guidelines with links.
2. The 2nd row of the table quotes guideline sections. For the newest guidelines (NICE, 2014), several quoted guideline statements contain the word "should", e.g., "1.3.1.3 Early intervention in psychosis services should aim to provide a full range of pharmacological, psychological, social, occupational and educational interventions for people with psychosis, consistent with this guideline. **[2014]."** No "must" statements are evident in this quoted sections.
3. The 2nd row of the table also quotes sections from practice guidelines prior to 2014, and those seem to be graded with a "[I]" indicator, e.g., "Treatment programs need to combine medications with a range of psychosocial services to reduce the need for crisis-oriented hospitalizations and emergency department visits and enable greater recovery [I]."
4. The third row of the table refers to methods used to grade evidence and cited Guyatt et al., 2011 and Mustafa et al., 2013. Higher grades are given to randomize trials than to observational studies, and raters are trained to judge studies, etc. This section, however, does not assign any grades to the 2014 NICE guidelines, and it says nothing about grades assigned to the 2014 guidelines.
5. The 4th row appears completely redundant with the third row. Again here, no grades are assigned to the recommendations. (This thus appears to be an error of some kind by the developer)
6. In the 4th row of the recommendation grading scheme seems to describe the range of grades for the 2014 guideless from highest ("must") to moderate ("should") to lowest ("could"). The developer here does not actually assign grades for the reader, but these statements along with the note 2 above suggests that the evidence they present from the guidelines never exceed a "should" (i.e., moderate) rating. For older evidence they do note that the rating achieved was "[I] Recommended with substantial clinical confidence."
7. In the 5th row the developer was asked to provide all other grades not assigned, but instead for the 2014 guidelines they repeat, verbatim, the three levels from the fourth row. For the older guidelines they list the other possible grades: "[II] Recommended with moderate clinical confidence. [III] May be recommended on the basis of individual circumstances."
8. In the 6th row they state the 2014 guidelines are "…based on the best available evidence" without describing anything about the quality or quantity of that evidence. They do not offer any body of evidence summary for the earlier guidelines.
9. In the 7th row they say the studies were reviewed by the GDG, but this paragraph seems to add little to the evidentiary presentation. Nothing was said about estimates of benefits and consistency regarding the earlier guidelines.
10. In the 8th row, they simply note for the 2014 guidelines that "no harms are cited." Harms were not addressed for the earlier guidelines.
11. In the 9th row, the last of the table, the developers state that >100 more recent studies have been published since the guidelines were published "none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness."

- **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

See description above and note that the developer seems to believe the existence of the guidelines and their methods is enough evidence, as the developers in their application do little to describe details about those guidelines and the evidence that lies beneath them.

Changes to evidence from last review

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

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

Updates:

Exception to evidence

[Exception to evidence]

Questions for the Committee:

- *Does the review of the three guidelines support the importance of the measure as being substantially linked to better outcomes which can be affected by the health care system?*

Guidance from the Evidence Algorithm

Not outcome→ Box 3 (Systematic review embedded in guidelines; 3 reviews cited) → Box 4 (QQC review is general, not discernably comprehensive) → Box 6 (suggested as "should" or "recommended with substantial clinical confidence") → moderate or low (committee discretion)

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

RATIONALE: Guideline support with three separate empirical studies supporting the principal of follow-up care.

Preliminary rating for evidence: ☐ **Pass** ☐ **No Pass** (Committee should discuss)

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.

Using well over 100 providers per sample, the following score-level distributions were proffered in section 1.b.2:

Commercial health plans, 30-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 60.1% | 11.5% | 43.9% | 52.3% | 60.4% | 67.1% | 75.0% | 14.8%

2016 | 61.3% | 11.5% | 45.8% | 54.7% | 62.1% | 68.6% | 75.8% | 13.9%

Commercial health plans, 7-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 45.3% | 12.3% | 30.3% | 37.7% | 44.4% | 51.8% | 62.0% | 14.1%

2016 | 45.8% | 11.5% | 31.1% | 38.6% | 45.2% | 53.1% | 60.4% | 14.6%

Medicaid health plans, 30-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 54.8% | 14.7% | 37.8% | 45.6% | 52.8% | 66.3% | 74.5% | 20.7%

2016 | 54.8% | 14.0% | 38.4% | 45.6% | 54.8% | 63.0% | 74.2% | 17.4%

Medicaid health plans, 7-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 40.1% | 15.4% | 21.9% | 28.9% | 37.3% | 53.0% | 61.3% | 24.1%

2016 | 39.5% | 14.6% | 23.0% | 29.8% | 37.5% | 47.4% | 61.1% | 17.6%

Medicare health plans, 30-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 47.6% | 13.5% | 31.3% | 37.8% | 47.0% | 56.6% | 65.0% | 18.8%

2016 | 49.5% | 14.0% | 32.9% | 41.2% | 49.1% | 59.1% | 66.1% | 17.9%

Medicare health plans, 7-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 31.5% | 13.6% | 16.3% | 21.7% | 29.0% | 39.1% | 49.6% | 17.4%

2016 | 33.4% | 13.3% | 17.6% | 23.7% | 32.0% | 43.0% | 51.7% | 19.2%

Disparities

See above sections differences between score-level rates for Commercial, Medicaid, and Medicare.

Questions for the Committee:

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

RATIONALE: Performance across hundreds of plans is well below 100% and demonstrates substantial spread between plans, and player types (Commercial, Medicare, Medicaid)

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

1a. Evidence

Comments:

**The relationship between the specific measure and the cited evidence is a bit loose (i.e., although many of the guidelines recommend some variant of continuity of care, they don't really translate into specific follow up times after ER visits). Having said that, follow up after these visits seems a self-evidently good idea

1b. Performance Gap

Comments:

**Yes, there's a gap in top end, median, and low end performance within and across payors.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers. For maintenance measures – less emphasis if no new testing data provided.

Validity

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

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

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

Evaluators: NQF Staff

Evaluation of Reliability and Validity:

- Reliability testing was conducted using and Adams-R score to assess whether between health plan variability presents as markedly greater than within plan variation. Adams-R did support such between plan variability.
- Validity testing was done empirically at the measure core lvel using a very similar measures (e.g., 30 day vs. 7 day; r >0.92) a related measures (alcohol/drug abuse follow-up; r >.42). A TEP was also consulted.

Questions for the Committee regarding reliability:

- What is the spread of the Adams-R calculations

Questions for the Committee regarding validity:

- Was the 7 to 30 day comparison really persuasive as a validity measure?
- Were the alcohol to mental health correlations strong enough to support validity?

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

Evaluation A: Scientific Acceptability

Evaluating Scientific Acceptability: Instructions

Scientific Acceptability: Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion.

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3489

Measure Title: Follow-Up After Emergency Department Visit for Mental Illness

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.)

- Measure was previously 2605: Follow-Up After Emergency Department Visit for Mental Illness or Alcohol and Other Drug Abuse or Dependence.
- The measure has been split into two separate measures:
  - 3489: Follow-Up After Emergency Department Visit for Mental Illness
  - 3488: Follow-Up After Emergency Department Visit for Alcohol and other Drug Dependence

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

2. **Briefly summarize any concerns about the measure specifications.**

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 <u>VALIDITY</u> testing** of <u>patient-level data</u> conducted?

   ☐ **Yes** ☐ **No** **n/a**

6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

Score-level method appears to be appropriate. Binary outcome (FU or not), scores tested with Adams-R (beta binomial model).

7. **Assess the results of reliability testing**

   Table 2. *Follow-Up After Emergency Department Visit for Mental Illness* Beta-Binomial Statistic, 2017

	Overall Reliability		
Measure Rate	Commercial	Medicaid	Medicare

30-day follow-up	0.95	0.99	0.91
7-day follow-up	0.96	0.99	0.92

The above results taken directly from the application suggest high reliability, but confidence interval or other spread information would strength that presentation. They should also note explicitly in their presentation that they, presumably, are presenting the mean R value across the sample of providers (which is over 166 per provider type (Commercial, Medicaid, Medicare)), per section 1.6.

Submission document: Testing attachment, section 2a2.3

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.**

    Rating might be higher if spread statistics had been presented, and if tables had clearly labeled the presented statistics as means.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Some concern that disenrollees are excluded from denominator because this means persons with coverage lapses may be completely ignored for needed follow-up. At least presenting statistics about how many such disenrollees are observed would be helpful. Quantifying both the simple magnitude and the mental health impact of the hospice exclusion could also be useful.

    Submission document: Testing attachment, section 2b2.

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

    No concerns

Submission document: Testing attachment, section 2b4.

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.

The developer says HEDIS reporting audits exist to "to verify primary data sources used to populate measures and ensure specifications are correctly implemented." The following question remain: do claims records overall contain the information needed for this measure, or is missing data a concern.

16. **Risk Adjustment** n/a

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

This involved the following:

1. Correlations between 30-day and 7-day scores,
2. Correlations between this measure and a similar ED follow-up measure for Alcohol and drug abuse,
3. OLD submission: Correlations of the inpatient use and the measure, with the a priori hypothesis that higher FU rates will correlate with lower inpatient rates, by state,
4. Face validity review by a multi-stakeholder advisor group.

20. **Assess the results(s) for establishing validity**

Submission document: Testing attachment, section 2b2.3

7-day versus 30-day correlation coefficients exceed 0.92 ($p<.05$), calculations made for each health plan type (Commercial, Medicaid, Medicare). This is not a very high bar to traverse.

Correlations between the alcohol and other drug follow-up measure and the current mental illness measure ranged from 0.42 to 0.57 ($p<.05$).

OLD submission- no significant difference was previously observed regarding inpatient mental illness hospitalizations in states with very high (top 25 percentile) versus very low (bottom 25th percentile FU rates (for alcohol, drug, or mental illness): 1.7% versus 2.1%, respectively ($p=.8$).

TEP and multi-stakeholder input was supportive of the measure, but the recency of this TEP is inclear and it seems to date back to a previous submission.

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 <u>are</u> threats to validity and/or relevant threats to validity were <u>not assessed OR</u> 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 <u>is required</u>; 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 done, but the comparators (i.e., external standards) were only somewhat persuasive that the measure is truly capturing mental illness follow-up services that yield better outcomes.

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

2a2. Reliability – Testing
Comments:
**No

2b1. Validity –Testing
Comments:
**No

2b2-3. Meaningful Differences
Comments:
**No concerns

2b4-7. Threats to Validity
Comments:
**No concerns

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.

Claims-based with NCQA history that includes auditing procedures to consider quality of inputs and outputs, presumably. Also, this measure is available for broad public use, though some commercial use and purchasing provisions may apply.

Questions for the Committee:

- Are there any concerns about commercial use provisions?
- Any concerns about claims data as a sole source for this measure?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. Feasibility
Comments:

**I wonder about only including people with a primary diagnosis of a mental health disorder. Is anything known about how often people with both a mental health disorder and some physical ailment are seen in ERs without a primary diagnosis. should there be consideration of over time expanding to people with either a primary or secondary mental health dx

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

https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html
CMS Medicaid Adult Core Set
NCQA Health Plan Ratings / Report Cards
https://www.ncqa.org/hedis/reports-and-research/ratings-methodology-and-guidelines/
Quality Improvement (external benchmarking to organizations)
NCQA Annual State of Health Care Quality
http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality
NCQA Quality Compass
http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass
SAMHSA Demonstration Program for Certified Community Behavioral Health Clinics (CCBHCs)
https://www.samhsa.gov/section-223

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

The developer says they share, disseminate, tender feedback, and incorporate that feedback into their measure. The best evidence of that are the numerous programs and health plans that use this measure.

Additional Feedback:

Questions for the Committee:

- The developer notes that performance rates on this measure over the past 2 years have not increased (i.e., remained steady), or have declined slightly in Medicare. How badly does this compromise the measure as a quality driver?

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

RATIONALE:

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

Over the two years studied (see section 1b above): the rates actually declined slightly in Medicare, though they were study in Commercial and Medicaid. The developer thus states: ". All health plans need to substantially improve follow-up care for mental health services. "

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

Unexpected findings (positive or negative) during implementation [unexpected findings]

None identified by developer

Potential harms [potential harms]

None identified by developer

Additional Feedback: none

Questions for the Committee:

- No problem report, but as noted above, improvement on the measure was not observed.

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

RATIONALE:

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

4a1. Use - Accountability and Transparency
Comments:

**Measure is being used and reported

4b1. Usability – Improvement
Comments:
**No substantial concerns

Criterion 5: Related and Competing Measures

Related or competing measures

0576 : Follow-Up After Hospitalization for Mental Illness (FUH)

Harmonization

"The following highlights the differences between the measures: Population focus (denominator): The measure targets patients discharged from the emergency department (not inpatient). Numerator: The measure captures follow-up with a primary mental health diagnosis (regardless of the type of provider)."

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

5. Related and Competing
Comments:
**No substantial concerns

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 06/17/2019

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

Brief Measure Information

NQF #: 3489

Corresponding Measures:

De.2. Measure Title: Follow-Up After Emergency Department Visit for Mental Illness

Co.1.1. Measure Steward: National Committee for Quality Assurance

De.3. Brief Description of Measure: The percentage of emergency department (ED) visits for members 6 years of age and older with a principal diagnosis of mental illness or intentional self-harm, who had a follow-up visit for mental illness. Two rates are reported:

- The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days).

- The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days).

1b.1. Developer Rationale: This measure targets individuals with mental health who are discharged to the community from the emergency department. These individuals may be particularly vulnerable to losing contact with the health care system. High use of the emergency department may signal a lack of access to ongoing care or a gap in fulfilling urgent care needs. Therefore, this point of transition presents an opportunity to ensure that the patient is connected to care and receives follow-up. Health plans have access to information and care management processes to ensure that follow-up care occurs. Therefore, health plans can help connect patients into outpatient care after emergency department use.

Estimates suggest that about half of psychiatric patients discharged from the emergency room transitioned successfully to outpatient care (Bruffaerts, 2005). Low-intensity interventions that can be applied widely are typically implemented at periods of high risk for treatment dropout, such as following an emergency room discharge or the time of entry into outpatient treatment (Kreyenbuhl, 2009).

Individuals discharged from the emergency department face two main risks: (1) disengagement from treatment and (2) readmission to the emergency department. Treatment disengagement is a problem because individuals with the most serious mental health problems or alcohol or drug use disorders may require ongoing support and counseling to live independently in the community. A retrospective chart study of 390 patients assessed the likelihood of patients presenting to the ED after engagement with a transitional psychiatric clinic post ED visit. Patients who had appointments in the clinic within 3 days of being discharged from the ED were more likely to be engaged in their post ED care and stayed in the community longer before returning to the ED (McCullumsith, 2015). Appropriate follow up after ED visit for mental health is needed to improve patient outcomes and treatment adherence.

Bruffaerts R, Sabbe M. Demyffenaere K. (2005) Predicting Community Tenure in Patients with Recurrent Utilization of a Psychiatric Emergency Service. Gen Hosp Psychiatry. 27:269-74.

Kreyenbuhl, J, Nossel, I, Dixon, L. (2009) Disengagement from Mental Health Treatment among Individuals with Schizophrenia and Strategies for Facilitating Connections to Care: A Review of the literature. Schizophrenia Bulletin. 35:696-703.

McCullumsmith, C., Clark, B., Blair, C., Cropsey, K., & Shelton, R. (2015). Rapid follow-up for patients after psychiatric crisis. Community mental health journal, 51(2), 139-144.

S.4. Numerator Statement: The numerator consists of two rates:

- 30-day follow-up: The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days).

- 7-day follow-up: The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days).

S.6. Denominator Statement: Emergency department (ED) visits for members 6 years of age and older with a principal diagnosis of mental illness or intentional self-harm on or between January 1 and December 1 of the measurement year.

S.8. Denominator Exclusions: Patients in hospice.

De.1. Measure Type: Process

S.17. Data Source: Claims

S.20. Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: **As 2506 (with SUD) 3/6/15** **Most Recent Endorsement Date:** **As 2506 (with SUD) 11/29/18**

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

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

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

1. Evidence 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

FUM_Evidence_Form.docx

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

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

Yes

1a. Evidence (subcriterion 1a)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3489

Measure Title: Follow-Up After Emergency Department Visit for Mental Illness

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

Date of Submission: <u>4/2/2019</u>

Instructions

- ***Complete 1a.1 and 1a.2 for all measures. If instrument-based measure, complete 1a.3.***
- ***Complete EITHER 1a.2, 1a.3 or 1a.4 as applicable for the type of measure and evidence.***

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

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

1a. Evidence to Support the Measure Focus

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

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

Notes

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

4. The preferred systems for grading the evidence are the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines and/or modified GRADE.

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

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

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

Outcome

☐ Outcome: Click here to name the health outcome

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

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

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

☒ Process: The percentage of ED visits for which members 6 years of age and older received follow-up after a qualifying diagnosis and event

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

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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

2019 Submission

Emergency department visit >>> Primary diagnosis of mental health >>> Discharge from the emergency room to the community>>> Patient had an outpatient visit, intensive outpatient visit, partial hospitalization within 7 and 30 day time period with any provider >>> Condition or disease management >>> Improvement in health outcome

2014 Submission

Emergency department visit >>> Primary diagnosis of mental health or alcohol and other drug use or dependence >>> Discharge from the emergency room to the community>>> Patient had an outpatient visit, intensive outpatient visit, partial hospitalization within 7 and 30 day time period with any provider >>> Condition or disease management >>> Improvement in health outcome

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

2019 Submission

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.

2019 Submission

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**	**2019 Submission** **Psychosis and schizophrenia in adults: treatment and management.** **2014** **National Collaborating Centre for Mental Health. Psychosis and schizophrenia in adults: prevention and management. London (UK): National Institute for Health and Care Excellence (NICE); 2014 Mar. 58 p. (NICE clinical guideline; no 178).** **https://www.nice.org.uk/guidance/cg178/resources/psychosis-and-schizophrenia-in-adults-prevention-and-management-35109758952133** **2014 Submission** **American Psychiatric Association: Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition** **URL: http://psychiatryonline.org/content.aspx?bookID=28&sectionID=1665359#46264** **Year: 2004** **Practice Guideline for the Treatment of Patients With Bipolar Disorder, Second Edition** **URL: http://psychiatryonline.org/content.aspx?bookid=28&sectionid=1669577** **Year: 2002** **Practice Guideline for the Treatment of Patients With Major Depressive Disorder, Third Edition** **URL: http://psychiatryonline.org/content.aspx?bookid=28&sectionid=1667485** **Year: 2010**
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being	**2019 Submission** **1.2 Preventing psychosis** 1.2.1 Referral from primary care 1.2.1.1 If a person is distressed, has a decline in social functioning and has: • transient or attenuated psychotic symptoms **or** • other experiences or behaviour suggestive of possible psychosis **or**

measured. If not a guideline, summarize the conclusions from the SR.	• a first-degree relative with psychosis or schizophrenia refer them for assessment without delay to a specialist mental health service or an early intervention in psychosis service because they may be at increased risk of developing psychosis. **[new 2014]** 1.2.2 Specialist assessment • 1.2.2.1 A consultant psychiatrist or a trained specialist with experience in at-risk mental states should carry out the assessment. **[new 2014]** ## 1.3 First episode psychosis ### 1.3.1 Early intervention in psychosis services • 1.3.1.3 Early intervention in psychosis services should aim to provide a full range of pharmacological, psychological, social, occupational and educational interventions for people with psychosis, consistent with this guideline. **[2014]** • 1.3.1.4 Consider extending the availability of early intervention in psychosis services beyond 3 years if the person has not made a stable recovery from psychosis or schizophrenia. **[new 2014]** ### 1.3.3 Assessment and care planning • 1.3.3.1 Carry out a comprehensive multidisciplinary assessment of people with psychotic symptoms in secondary care. This should include assessment by a psychiatrist, a psychologist or a professional with expertise in the psychological treatment of people with psychosis or schizophrenia. ### 1.4.6 Early post-acute period • 1.4.6.1 After each acute episode, encourage people with psychosis or schizophrenia to write an account of their illness in their notes. **[2009]** • 1.4.6.2 Healthcare professionals may consider using psychoanalytic and psychodynamic principles to help them understand the experiences of people with psychosis or schizophrenia and their interpersonal relationships. **[2009]** • 1.4.6.3 Inform the service user that there is a high risk of relapse if they stop medication in the next 1–2 years. **[2009]** • 1.4.6.4 If withdrawing antipsychotic medication, undertake gradually and monitor regularly for signs and symptoms of relapse. **[2009]** 1.4.6.5 After withdrawal from antipsychotic medication, continue monitoring for signs and symptoms of relapse for at least 2 years. **[2009]** **2014 Submission** Continuity of care is an important issue for individuals with mental illness. Existing clinical practice guidelines recommend ongoing monitoring and management using a variety of interventions, but do not explicitly address post emergency department follow-up.

	American Psychiatric Association: Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition, Page 14, "Treatment programs need to combine medications with a range of psychosocial services to reduce the need for crisis-oriented hospitalizations and emergency department visits and enable greater recovery [I]." Practice Guideline for the Treatment of Patients With Bipolar Disorder Second Edition, Page 9, "Subsequently, specific goals of psychiatric management include establishing and maintaining a therapeutic alliance, monitoring the patient's psychiatric status, providing education regarding bipolar disorder, enhancing treatment compliance, promoting regular patterns of activity and of sleep, anticipating stressors, identifying new episodes early, and minimizing functional impairments [I]." Practice Guideline for the Treatment of Patients With Major Depressive Disorder, Third Edition, Page 15, "Psychiatric management consists of a broad array of interventions and activities that psychiatrists should initiate and continue to provide to patients with major depressive disorder through all phases of treatment [I]."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2019 Submission** For questions about the effectiveness of interventions, the GRADE approach was used to grade the quality of evidence for each outcome (Guyatt et al., 2011). For questions about the experience of care and the organisation and delivery of care, methodology checklists (see section 3.5.1) were used to assess the risk of bias, and this information was taken into account when interpreting the evidence. The technical team produced GRADE evidence profiles (see below) using GRADE profiler (GRADEpro) software (Version 3.6), following advice set out in the GRADE handbook (Schünemann et al., 2009). Those doing GRADE ratings were trained, and calibration exercises were used to improve reliability (Mustafa et al., 2013). A GRADE evidence profile was used to summarise both the quality of the evidence and the results of the evidence synthesis for each 'critical' and 'important' outcome. The GRADE approach is based on a sequential assessment of the quality of evidence, followed by judgment about the balance between desirable and undesirable effects, and subsequent decision about the strength of a recommendation. Within the GRADE approach to grading the quality of evidence, the following is used as a starting point: • RCTs without important limitations provide high quality evidence • observational studies without special strengths or important limitations provide low quality evidence. For each outcome, quality may be reduced depending on five factors: methodological limitations, inconsistency, indirectness, imprecision and publication bias. For the purposes of the guideline, each factor was evaluated using criteria provided in Table 4. For observational studies without any reasons for down-grading, the quality may be up-graded if there is a large effect, all plausible confounding would reduce the demonstrated effect (or increase the effect if no effect was observed), or there is evidence of a dose-response gradient (details would be provided under the 'other' column). Each

	evidence profile includes a summary of findings: number of participants included in each group, an estimate of the magnitude of the effect, and the overall quality of the evidence for each outcome. Under the GRADE approach, the overall quality for each outcome is categorised into one of four groups (high, moderate, low, very low). https://www.nice.org.uk/guidance/cg178/evidence/appendix-13-490503567 **2014 Submission**
Provide all other grades and definitions from the evidence grading system	**2019 Submission** For questions about the effectiveness of interventions, the GRADE approach was used to grade the quality of evidence for each outcome (Guyatt et al., 2011). For questions about the experience of care and the organisation and delivery of care, methodology checklists (see section 3.5.1) were used to assess the risk of bias, and this information was taken into account when interpreting the evidence. The technical team produced GRADE evidence profiles (see below) using GRADE profiler (GRADEpro) software (Version 3.6), following advice set out in the GRADE handbook (Schünemann et al., 2009). Those doing GRADE ratings were trained, and calibration exercises were used to improve reliability (Mustafa et al., 2013). A GRADE evidence profile was used to summarise both the quality of the evidence and the results of the evidence synthesis for each 'critical' and 'important' outcome. The GRADE approach is based on a sequential assessment of the quality of evidence, followed by judgment about the balance between desirable and undesirable effects, and subsequent decision about the strength of a recommendation. Within the GRADE approach to grading the quality of evidence, the following is used as a starting point: • RCTs without important limitations provide high quality evidence • observational studies without special strengths or important limitations provide low quality evidence. For each outcome, quality may be reduced depending on five factors: methodological limitations, inconsistency, indirectness, imprecision and publication bias. For the purposes of the guideline, each factor was evaluated using criteria provided in Table 4. For observational studies without any reasons for down-grading, the quality may be up-graded if there is a large effect, all plausible confounding would reduce the demonstrated effect (or increase the effect if no effect was observed), or there is evidence of a dose-response gradient (details would be provided under the 'other' column). Each evidence profile includes a summary of findings: number of participants included in each group, an estimate of the magnitude of the effect, and the overall quality of the evidence for each outcome. Under the GRADE approach, the overall quality for each outcome is categorised into one of four groups (high, moderate, low, very low). https://www.nice.org.uk/guidance/cg178/evidence/appendix-13-490503567 **2014 Submission**

Grade assigned to the **recommendation** with definition of the grade	**2019 Submission** The description of the process of moving from evidence to recommendations indicates that some recommendations can be made with more certainty than others. This concept of the 'strength' of a recommendation should be reflected in the consistent wording of recommendations within and across clinical guidelines. There are three levels of certainty: • recommendations for interventions that must (or must not) be used: Recommendations that an intervention must or must not be used are usually included only if there is a legal duty to apply the recommendation, for example to comply with health and safety regulations. In these instances, give a reference to supporting documents. These recommendations apply to all patients. • recommendations for interventions that should (or should not) be used: For recommendations on interventions that 'should' be used, the GDG is confident that, for the vast majority of people, the intervention (or interventions) will do more good than harm, and will be cost effective. • recommendations for interventions that could be used: For recommendations on interventions that 'could' be used, the GDG is confident that the intervention will do more good than harm for most patients, and will be cost effective Recommendations are marked as **[2009]**, **[2009, amended 2014]**, **[2014]** or **[new 2014]**. • **[2009]** indicates that the evidence has not been reviewed since 2009. • **[2009, amended 2014]** indicates that the evidence has not been reviewed since 2009 but changes have been made to the recommendation wording that change the meaning. • **[2014]** indicates that the evidence has been reviewed but no changes have been made to the recommendation. **[new 2014]** indicates that the evidence has been reviewed and the recommendation has been updated or added. **2014 Submission** [I] Recommended with substantial clinical confidence.
Provide all other grades and definitions from the recommendation grading system	**2019 Submission** The description of the process of moving from evidence to recommendations indicates that some recommendations can be made with more certainty than others. This concept of the 'strength' of a recommendation should be reflected in the consistent wording of recommendations within and across clinical guidelines. There are three levels of certainty: • recommendations for interventions that must (or must not) be used: Recommendations that an intervention must or must not be used are usually included only if there is a legal duty to apply the recommendation, for example to comply with health and safety

	regulations. In these instances, give a reference to supporting documents. These recommendations apply to all patients. • recommendations for interventions that should (or should not) be used: For recommendations on interventions that 'should' be used, the GDG is confident that, for the vast majority of people, the intervention (or interventions) will do more good than harm, and will be cost effective. • recommendations for interventions that could be used: For recommendations on interventions that 'could' be used, the GDG is confident that the intervention will do more good than harm for most patients, and will be cost effective Recommendations are marked as **[2009]**, **[2009, amended 2014]**, **[2014]** or **[new 2014]**. • **[2009]** indicates that the evidence has not been reviewed since 2009. • **[2009, amended 2014]** indicates that the evidence has not been reviewed since 2009 but changes have been made to the recommendation wording that change the meaning. • **[2014]** indicates that the evidence has been reviewed but no changes have been made to the recommendation. • **[new 2014]** indicates that the evidence has been reviewed and the recommendation has been updated or added. **2014 Submission** [II] Recommended with moderate clinical confidence. [III] May be recommended on the basis of individual circumstances.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	NICE guideline recommendations are based on the best available evidence. We use a wide range of different types of evidence and other information – from scientific research using a variety of methods, to testimony from practitioners and people using services. **2014 Submission**
Estimates of benefit and consistency across studies	All primary-level studies included after the first scan of citations were acquired in full and re-evaluated for eligibility at the time they were being entered into the study information database. More specific eligibility criteria were developed for each review question and are described in the relevant clinical evidence chapters. Eligible systematic reviews and primary-level studies were critically appraised for methodological quality (risk of bias) using a checklist (see The Guidelines Manual (NICE, 2012b) for templates). The eligibility of each study was confirmed by at least one member of the GDG. **2014 Submission**
What harms were identified?	No identified harms are cited.

	2014 Submission
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness. **2014 Submission**

_____

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.

2019 Submission
Not applicable

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

2019 Submission
Not applicable

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

2019 Submission
Not applicable

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

2019 Submission
Not applicable

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 targets individuals with mental health who are discharged to the community from the emergency department. These individuals may be particularly vulnerable to losing contact with the health care system. High use of the emergency department may signal a lack of access to ongoing care or a gap in fulfilling urgent care needs. Therefore, this point of transition presents an opportunity to ensure that the patient is connected

to care and receives follow-up. Health plans have access to information and care management processes to ensure that follow-up care occurs. Therefore, health plans can help connect patients into outpatient care after emergency department use.

Estimates suggest that about half of psychiatric patients discharged from the emergency room transitioned successfully to outpatient care (Bruffaerts, 2005). Low-intensity interventions that can be applied widely are typically implemented at periods of high risk for treatment dropout, such as following an emergency room discharge or the time of entry into outpatient treatment (Kreyenbuhl, 2009).

Individuals discharged from the emergency department face two main risks: (1) disengagement from treatment and (2) readmission to the emergency department. Treatment disengagement is a problem because individuals with the most serious mental health problems or alcohol or drug use disorders may require ongoing support and counseling to live independently in the community. A retrospective chart study of 390 patients assessed the likelihood of patients presenting to the ED after engagement with a transitional psychiatric clinic post ED visit. Patients who had appointments in the clinic within 3 days of being discharged from the ED were more likely to be engaged in their post ED care and stayed in the community longer before returning to the ED (McCullumsith, 2015). Appropriate follow up after ED visit for mental health is needed to improve patient outcomes and treatment adherence.

Bruffaerts R, Sabbe M. Demyffenaere K. (2005) Predicting Community Tenure in Patients with Recurrent Utilization of a Psychiatric Emergency Service. Gen Hosp Psychiatry. 27:269-74.

Kreyenbuhl, J, Nossel, I, Dixon, L. (2009) Disengagement from Mental Health Treatment among Individuals with Schizophrenia and Strategies for Facilitating Connections to Care: A Review of the literature. Schizophrenia Bulletin. 35:696-703.

McCullumsmith, C., Clark, B., Blair, C., Cropsey, K., & Shelton, R. (2015). Rapid follow-up for patients after psychiatric crisis. Community mental health journal, 51(2), 139-144.

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

The following data are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. Performance data is summarized at the health plan level and summarized by mean, standard deviation, and performance at 10th, 25th, 50th, 75th, and 90th percentile. Data is stratified by year and product line (i.e. commercial, Medicare, Medicaid).

Commercial health plans, 30-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 60.1% | 11.5% | 43.9% | 52.3% | 60.4% | 67.1% | 75.0% | 14.8%

2016 | 61.3% | 11.5% | 45.8% | 54.7% | 62.1% | 68.6% | 75.8% | 13.9%

Commercial health plans, 7-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 45.3% | 12.3% | 30.3% | 37.7% | 44.4% | 51.8% | 62.0% | 14.1%

2016 | 45.8% | 11.5% | 31.1% | 38.6% | 45.2% | 53.1% | 60.4% | 14.6%

Medicaid health plans, 30-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 54.8% | 14.7% | 37.8% | 45.6% | 52.8% | 66.3% | 74.5% | 20.7%

2016 | 54.8% | 14.0% | 38.4% | 45.6% | 54.8% | 63.0% | 74.2% | 17.4%

Medicaid health plans, 7-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 40.1% | 15.4% | 21.9% | 28.9% | 37.3% | 53.0% | 61.3% | 24.1%

2016 | 39.5% | 14.6% | 23.0% | 29.8% | 37.5% | 47.4% | 61.1% | 17.6%

Medicare health plans, 30-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 47.6% | 13.5% | 31.3% | 37.8% | 47.0% | 56.6% | 65.0% | 18.8%

2016 | 49.5% | 14.0% | 32.9% | 41.2% | 49.1% | 59.1% | 66.1% | 17.9%

Medicare health plans, 7-day follow-up

YEAR| MEAN | ST DEV | 10TH | 25TH | 50TH | 75TH | 90TH | Interquartile Range

2017 | 31.5% | 13.6% | 16.3% | 21.7% | 29.0% | 39.1% | 49.6% | 17.4%

2016 | 33.4% | 13.3% | 17.6% | 23.7% | 32.0% | 43.0% | 51.7% | 19.2%

The data references are extracted from HEDIS data collection reflecting the most recent years of measurement for this measure. In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries and 47.0 million Medicaid beneficiaries. Below is a description of the denominator for this measure. It includes the number of health plans included in HEDIS data collection and the median eligible number of encounters for the measure across health plans.

Commercial health plans, 30-day follow-up

2017 | 319 | 194

2016 | 307 | 193

Commercial health plans, 7-day follow-up

2017 | 319 | 194

2016 | 307 | 193

Medicaid health plans, 30-day follow-up

2017 | 166 | 698

2016 | 128 | 650

Medicaid health plans, 7-day follow-up

2017 | 166 | 698

2016 | 128 | 650

Medicare health plans, 30-day follow-up

2017 | 264 | 94

2016 | 245 | 117

Medicare health plans, 7-day follow-up

2017 | 264 | 94

2016 | 245 | 117

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.

There is room for improvement in measure performance. Average performance rates for both 7-day and 30-day follow up are low for commercial, Medicare, and Medicaid health plans. Average 30-day follow-up performance across all ages is 48 percent for Medicare plans, 55 percent for Medicaid plans, and 60 percent for

commercial plans. Average 7-day follow-up performance across all ages is 32 percent for Medicare plans, 40 percent for Medicaid plans, and 45 percent for commercial plans.

There is also a wide range in performance for both the 7-day and 30-day follow-up rates. For example, in 2017, Medicare plan performance for 7-day follow-up ranged from 16 percent (plans in the 10th percentile) to 50 percent (plans in the 90th percentile). 30-day follow-up rates similarly show a wide range; for example, commercial health plan performance ranged from 44 percent (in the 10th percentile) to 75 percent (in the 90th percentile).

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

HEDIS data are stratified by type of insurance (e.g. Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities, if the data are available to a plan. The HEDIS Race/Ethnicity Diversity of Membership and the Language Diversity of Membership measures were designed to promote standardized methods for collecting these data and follow Office of Management and Budget and Institute of Medicine guidelines for collecting and categorizing race/ethnicity and language data. In addition, NCQA's Multicultural Health Care Distinction Program outlines standards for collecting, storing, and using race/ethnicity and language data to assess health care disparities.

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

The highest rates of ED visits for psychiatric crisis are seen among African Americans, persons with Medicaid and the uninsured. For these patients, a lack of access to or poor engagement with outpatient psychiatric services might significantly contribute to ED visits for psychiatric crisis (Hazlett et al., 2004).

For patients with both depression and a substance use disorder, women are more likely than men to receive treatment (Satre et al, 2010).

Hazlett, S.B., M.L. McCarthy, M.S. Londner, & C.U. Onyike. 2004. "Epidemiology of adult psychiatric visits to US Emergency Departments." Academic Emergency Medicine 11(2), 193–195.

Satre, D., C.I. Campbell, N.P. Gordon, C. Weisner. "Ethnic disparities in accessing treatment for depression and substance use disorders in an integrated health plan." Int J Psychiatry Med. 2010 ; 40(1): 57–76.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the sub criteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

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

De.5. Subject/Topic Area *(check all the areas that apply):*

Behavioral Health, Behavioral Health : Alcohol, Substance Use/Abuse

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

Access to Care, Care Coordination, Disparities Sensitive

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

Populations at Risk

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

Not applicable.

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

This is not an eMeasure **Attachment:**

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

Attachment **Attachment:** 3489_FUM_Value_Sets_Spring_2019.xlsx

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

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

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

Not an instrument-based measure

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

Yes

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

Measure #2605, Follow-Up After Emergency Department Visit for Mental Illness or Alcohol and Other Drug Abuse or Dependence, has been split into two separate measures:

- Follow-Up After Emergency Department Visit for Mental Illness (#3489)

- Follow-Up After Emergency Department Visit for Alcohol and other Drug Dependence (#3488)

Added telehealth to the measure numerators.

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

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

The numerator consists of two rates:

- 30-day follow-up: The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days).

- 7-day follow-up: The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days).

S.5. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

30-day follow-up: The percentage of ED visits for which the member received follow-up within 30 days of the ED visit (31 total days). Any of the following meet criteria for a follow-up visit:

- An outpatient visit (Visit Setting Unspecified Value Set with Outpatient POS Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An outpatient visit (BH Outpatient Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Visit Setting Unspecified Value Set with Partial Hospitalization POS Value Set), with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Partial Hospitalization/Intensive Outpatient Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A community mental health center visit (Visit Setting Unspecified Value Set with Community Mental Health Center POS Value Set), with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- Electroconvulsive therapy (Electroconvulsive Therapy Value Set) with (Ambulatory Surgical Center POS Value Set; Community Mental Health Center POS Value Set; Outpatient POS Value Set; Partial Hospitalization POS Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A telehealth visit (Visit Setting Unspecified Value Set with Telehealth POS Value Set), with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An observation visit (Observation Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- An outpatient visit (Visit Setting Unspecified Value Set with Outpatient POS Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An outpatient visit (BH Outpatient Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Visit Setting Unspecified Value Set with Partial Hospitalization POS Value Set), with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Partial Hospitalization/Intensive Outpatient Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A community mental health center visit (Visit Setting Unspecified Value Set with Community Mental Health Center POS Value Set), with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set),

with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- Electroconvulsive therapy (Electroconvulsive Therapy Value Set) with (Ambulatory Surgical Center POS Value Set; Community Mental Health Center POS Value Set; Outpatient POS Value Set; Partial Hospitalization POS Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A telehealth visit (Visit Setting Unspecified Value Set with Telehealth POS Value Set), with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An observation visit (Observation Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

7-day follow-up: The percentage of ED visits for which the member received follow-up within 7 days of the ED visit (8 total days). Any of the following meet criteria for a follow-up visit:

- An outpatient visit (Visit Setting Unspecified Value Set with Outpatient POS Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An outpatient visit (BH Outpatient Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Visit Setting Unspecified Value Set with Partial Hospitalization POS Value Set), with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Partial Hospitalization/Intensive Outpatient Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A community mental health center visit (Visit Setting Unspecified Value Set with Community Mental Health Center POS Value Set), with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- Electroconvulsive therapy (Electroconvulsive Therapy Value Set) with (Ambulatory Surgical Center POS Value Set; Community Mental Health Center POS Value Set; Outpatient POS Value Set; Partial Hospitalization POS Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A telehealth visit (Visit Setting Unspecified Value Set with Telehealth POS Value Set), with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An observation visit (Observation Value Set) with a principal diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- An outpatient visit (Visit Setting Unspecified Value Set with Outpatient POS Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An outpatient visit (BH Outpatient Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Visit Setting Unspecified Value Set with Partial Hospitalization POS Value Set), with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An intensive outpatient encounter or partial hospitalization (Partial Hospitalization/Intensive Outpatient Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A community mental health center visit (Visit Setting Unspecified Value Set with Community Mental Health Center POS Value Set), with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- Electroconvulsive therapy (Electroconvulsive Therapy Value Set) with (Ambulatory Surgical Center POS Value Set; Community Mental Health Center POS Value Set; Outpatient POS Value Set; Partial Hospitalization POS Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

- A telehealth visit (Visit Setting Unspecified Value Set with Telehealth POS Value Set), with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set), with or without a telehealth modifier (Telehealth Modifier Value Set).

- An observation visit (Observation Value Set) with a principal diagnosis of intentional self-harm (Intentional Self-Harm Value Set), with any diagnosis of a mental health disorder (Mental Health Diagnosis Value Set).

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

Emergency department (ED) visits for members 6 years of age and older with a principal diagnosis of mental illness or intentional self-harm on or between January 1 and December 1 of the measurement year.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

IF an OUTCOME MEASURE, describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).

Age: 6 years and older as of the date of the ED visit

Benefit: Medical and mental health.

Continuous Enrollment: Date of emergency department visit through 30 days the ED visit

Event/diagnosis criteria: An ED visit (ED Value Set) with a principal diagnosis of mental illness or intentional self-harm (Mental Illness Value Set; Intentional Self-Harm Value Set) on or between January 1 and December 1 of the measurement year where the member was 6 years or older on the date of the visit.

The denominator for this measure is based on ED visits, not on members. If a member has more than one ED visit, identify all eligible ED visits between January 1 and December 1 of the measurement year and do not include more than one visit per 31-day period as described below.

If a member has more than one ED visit in a 31-day period, include only the first eligible ED visit. For example, if a member has an ED visit on January 1 then include the January 1 visit and do not include ED visits that occur on or between January 2 and January 31; then, if applicable, include the next ED visit that occurs on or after February 1. Identify visits chronologically including only one per 31-day period. Note: Removal of multiple visits in a 31-day period is based on eligible visits. Assess each ED visit for exclusions before removing multiple visits in a 31-day period.

Exclude ED visits that result in an inpatient stay and ED visits followed by admission to an acute or nonacute inpatient care setting on the date of the ED visit or within the 30 days after the ED visit (31 total days), regardless of principal diagnosis for the admission. To identify admissions to an acute or nonacute inpatient care setting:

1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).

2. Identify the admission date for the stay.

These events are excluded from the measure because admission to an acute or nonacute inpatient setting may prevent an outpatient follow-up visit from taking place.

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

Patients in hospice.

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

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

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

Not applicable.

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

Step 1: Determine the eligible population.

Step 1A: Identify patients with who were treated and discharged from an emergency department with a primary diagnosis of mental health. Do not include ED visits that result in an inpatient stay, or are followed by an admission to an acute or nonacute inpatient care setting on the date of the ED visit or within the 30 days after the ED visit.

Step 2: Identify the numerator.

Step 2A: Identify those who had a qualifying follow-up visit within 7 days.

Step 2B: Identify those who had a qualifying follow-up visit within 30 days.

Step 3: Calculate the rates.

Step 3A: Calculate the 7-day rate by dividing the number of ED visits with qualifying follow-up visit within 7 days (Step 2A) by the denominator (Step 1A).

Step 3B: Calculate the 30-day rate by dividing the number of ED visits with qualifying follow-up visit within 30 days (Step 2B) by the denominator (Step 1A).

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

Not applicable.

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

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

Not applicable.

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

If other, please describe in S.18.

Claims

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g. name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

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

This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health Management Organizations and Preferred Provider Organizations via NCQA's online data submission system.

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

No data collection instrument provided

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

Health Plan

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

Outpatient Services

If other:

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

Not applicable.

2. Validity – See attached Measure Testing Submission Form

FUM_Measure_Testing_Form_April_2019.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)

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

Measure Number (*if previously endorsed*)**:** 3489
Measure Title: Follow-Up After Emergency Department Visit for Mental Illness
Date of Submission: 4/2/2019
Type of Measure:

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

Instructions

- **Measures must be tested for all the data sources and levels of analyses that are specified. *If there is more than one set of data specifications or more than one level of analysis, contact NQF staff* about how to present all the testing information in one form.**
- **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.**
- **For outcome and resource use measures, section 2b3 also must be completed.**
- **If specified for multiple data sources/sets of specificaitons (e.g., claims and EHRs), section 2b5 also must be completed.**
- **Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.**
- **If you are unable to check a box, please highlight or shade the box for your response.**
- **Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). *Contact NQF staff if more pages are needed.***
- **Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.**
- **For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment.**

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

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

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

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b3. For outcome measures and other measures when indicated (e.g., resource use):
- **an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration**

OR
- **rationale/data support no risk adjustment/ stratification.**

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful [16] differences in performance;
OR
there is evidence of overall less-than-optimal performance.

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

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

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation

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

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

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

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

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

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

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

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

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

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

2019 Submission

N/A

2014 Submission

Medicaid claims; Medicaid Analytic eXtract (MAX)

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

2019 Submission

Testing of measure score reliability and validity was performed using data from 2017.

2014 Submission

Calendar year 2008

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

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

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

2019 Submission

This measure assesses whether patients age 6 and older with an emergency department (ED) visit and a principal diagnosis of mental illness or intentional self-harm had a follow up visit for mental illness. This measure includes patients who were enrolled in commercial, Medicaid and Medicare health plans. There is a rate for the proportion of ED visits for which the patient received follow-up within 30 days of the ED visit, and a rate for the proportion of ED visits for which the patient received follow-up within 7 days of the ED visit. The intended use of the measure is to assess the quality of care in health plans across the population. As required by the specified level of accountability, we conducted a field test with health plans to assess scientific acceptability, usability and feasibility and have subsequently gathered audited data from a large number of health plans.

Sample for measure score reliability testing and construct validity testing: The measure score reliability was calculated from HEDIS data that included 319 commercial health plans, 166 Medicaid health plans, and 264 Medicare health plans. The sample included all health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

2014 Submission

RELIABILITY, VALIDITY AND MEANINGFUL DIFFERENCES

We tested the reliability, validity, and variation in performance on this measure among 16 states for the rate of follow-up for mental health (MH) emergency department visits and 15 states for the rate of follow-up for alcohol and other drug dependence (AOD) emergency department visits using fee-for-service (FFS) Medicaid claims derived from the MAX data. We used FFS claims because Medicaid managed care organizations do not submit encounters in many states or submit incomplete data that limits the ability to observe every medical or behavioral health encounter.

We excluded states where FFS data were not expected to be representative (e.g. where only a small percentage of Medicaid adults were enrolled in FFS), where there was a problem with the Medicaid enrollment file or with FFS claims (e.g. inability to identify our population of interest, or missing claims), or where the denominator size of emergency department discharges was very small (less than 150).

Systematic Evaluation of Face Validity

This measure was tested for validity with an expert panel (n=16), focus group (n=29), and public comment (n=20).

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*)

2019 Submission

Patient sample for measure score reliability testing: Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Table 1. Median denominator size for the *Follow-Up After Emergency Department Visit for Mental Illness* measure by plan type, 2017

Product Type	Number of Plans	Median number of encounters per plan
Commercial	319	194
Medicaid	166	698
Medicare	264	94

2014 Submission

Our analysis includes all Medicaid enrollees ages 18 and over. We excluded enrollees for whom Medicaid data would not be expected to include all instances of care provision including individuals who were (1) dually eligible for Medicare, (2) did not have full Medicaid benefits, (3) had private insurance, or (4) were enrolled in Medicaid for less than one calendar year.

The measure is calculated for two populations: (1) patients with a mental health emergency department visit and (2) patients with an alcohol or other drug dependence emergency department visit. For each population, there are two rates – follow-up within 7 days of emergency department discharge and follow up within 30 days of emergency department discharge. Table 1 summarizes the number and characteristics of individuals used to calculate the rates.

Table 1. Characteristics of patients in each denominator across all states included in analysis:

	Mental Health Denominator		AOD Denominator	
Number of states	N = 16		N = 15	
Characteristic	Number	Percentage	Number	Percentage
Total Individuals	26,982	100	11,743	100
Gender				
Male	10,744	39.8	6,068	51.7
Female	16,238	60.2	5,675	48.3
Unknown	0	0.0	0	0.0
Age				
15 to 20	2,015	7.5	550	4.7
21 to 44	15,602	57.8	5,447	46.4
45 to 64	9,214	34.1	5,656	48.2
65 to 74	132	0.5	84	0.7
75 to 84	17	0.1	6	0.1
85+	2	0.0	0	0.0
Race/Ethnicity				
African American	8,920	33.1	3,324	28.3
Caucasian	15,144	56.1	6,934	59.0
Hispanic	883	3.3	326	2.8
Other	485	1.8	377	3.2
Unknown	1,550	5.7	782	6.7
Medicaid Eligibility category				
Adult	3,877	14.4	1,876	16.0
Disabled	22,439	83.2	9,575	81.5
Children	666	2.5	292	2.5
Geography				
Metropolitan	11,146	41.3	5,021	42.8
Micropolitan	7,887	29.2	3,315	28.2

Neither	7,845	29.1	3,383	28.8
Unknown	104	0.4	24	0.2

Source: MAX data from calendar year 2008

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.

2019 Submission

No differences in the data used for reliability and construct validity testing.

2014 Submission

The number of states used for each denominator is different; 16 states were included in our analysis of the follow-up rate for emergency department visits for mental health diagnoses whereas 15 states were included in our analysis of the follow-up rate for emergency department visits for AOD diagnoses. As seen in Table 2, The District of Columbia was not included in the AOD analysis due to a small sample size. There were no other differences in the data used for each aspect of testing.

Table 2: Number of emergency department discharges included in each denominator, by state:

State	Number of ED discharges in Mental Health Denominator	Number of ED discharges in AOD Denominator
AK	221	212
AL	2,294	873
CT	1,608	1,135
DC*	181	N/A
GA	3,506	1,273
IL	5,681	1,248
IN	990	563
KY	3,520	1,403
LA	2,447	1,081
MN	2,149	747
MS	842	392
NC	4,907	2,416
NH	574	188
OK	813	514
WI	1,041	588
WV	1,178	704
Total	31,952	13,337

*DC was dropped from AOD denominator due to small sample size.

Source: MAX calendar year 2008

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.

2019 Submission

Social risk factor data were not available in reported results. This measure is specified to be reported separately by Medicare, Medicaid and commercial plan types, which serves as a proxy for income and other socioeconomic factors.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)
2019 Submission
Reliability testing of performance measure score
We utilized the Beta-binomial model (Adams 2009) to assess how well one can confidently distinguish the performance of one accountable entity from another. Conceptually, the Beta-binomial model is the ratio of signal to noise. The signal is the proportion of the variability in measured performance that can be explained by real differences in performance. The Beta-binomial model is an appropriate model when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS measures. Reliability scores range from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (i.e., noise), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

Adams, J.L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

2014 Submission
Reliability Testing of Performance Measure Score: In order to assess measure precision in the context of the observed variability across accountable entities, we used the beta-binomial method and resulting estimate described by Adams (2009). The following is quoted from the tutorial: "Reliability describes how well one can confidently distinguish the performance of one physician [or accountable entity] from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance." This approach is also relevant to health plans, states, and other accountable entities.

Adams' approach uses a beta-binomial model to estimate reliability; this model is suited for estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach assumes that the performance measure score (pass/fail rate) across accountable entities has a flexible beta distribution, characterized by a signal variance. Given its performance measure score, the observed data (number of passes/failures) for an accountable entity has a binomial distribution, which provides the noise (measurement error) variance. From the beta-binomial model, the signal and noise variances are used to calculate reliability as:

Signal variance / (signal + noise variance)

Reliability scores vary from 0.0 to 1.0. A score of zero indicates that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 indicates that all variation is attributable to real differences in performance across accountable entities.

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

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

2019 Submission

Table 2 shows the reliability for each indicator of measure as shown by the beta-binomial model.

Table 2. *Follow-Up After Emergency Department Visit for Mental Illness* Beta-Binomial Statistic, 2017

	Overall Reliability		
Measure Rate	Commercial	Medicaid	Medicare
30-day follow-up	0.95	0.99	0.91
7-day follow-up	0.96	0.99	0.92

2014 Submission

Reliability statistic for follow-up for MH emergency department visits:
Average, 7-day follow-up: .99
10^{th}-90^{th} percentile across states: .98 – 1.0

Average, 30-day follow-up: .98
10^{th}-90^{th} percentile across states: .97 – 1.0

Reliability statistic for follow-up for AOD emergency department visits:
Average, 7-day follow-up: .99
10^{th}-90^{th} percentile across states: .99 – 1.0

Average, 30-day follow-u: .99
10^{th}-90^{th} percentile across states: .98 – 1.0

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?*)

2019 Submission

In general, a score of 0.7 or higher suggests the measure has adequate reliability. The results suggest the measure has high reliability.

2014 Submission

Reliability Testing of Performance Measure Score: Reliability scores can vary from 0.0 to 1.0. Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The testing suggests the all four follow-up rates reported as part of this measure have strong reliability between .98 and .99.

The minimum state-level reliability scores for this measure all exceed the minimally accepted threshold of 0.7.

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)*

2019 Submission

We assessed face validity and construct validity for this measure.

Method of testing construct validity

We tested for construct validity by exploring the following:

- Are the individual rates within the *Follow-Up After Emergency Department Visit for Mental Illness* measure correlated with one another
- Is *Follow-Up After Emergency Department Visit for Mental Illness* correlated with the HEDIS *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* measure, which assesses the proportion of ED visits for alcohol or other drug abuse or dependence that had a follow-up visit

We hypothesized that rates within the *Follow-Up After Emergency Department Visit for Mental Illness* measure would be highly correlated, and that organizations that perform well on *Follow-Up After Emergency Department Visit for Mental Illness* should perform well on the other measure as they address the same or similar behavioral health conditions. To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two variables. The magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

Method of assessing face validity

NCQA develops measures using a standardized process. For new measures, face validity is assessed at various steps as described below.

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

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. At this step, face validity is systematically determined by the CPM, which uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment. For the most recent updates to this measure in January 2016, the CPM voted to approve moving the proposed changes forward to public comment (9 CPM members approved, 0 members opposed and 0 abstained).

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA about proposed new measures. Public comment offers an opportunity to assess the validity, feasibility, importance and other attributes of a measure from a wider audience. For this measure, a majority of public comment respondents supported the measure. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. Face validity is then again systematically assessed by the CPM. The CPM reviews all comments before making a final decision and votes to recommend approval of new measures for HEDIS. NCQA's Board of Directors then approves new measures.

For the most recent updates to this measure in May 2016, the CPM voted to approve the measure for HEDIS health plan reporting (13 CPM members approved, 0 members opposed and 0 abstained).

2014 Submission

Empirical validity testing

We tested for construct validity by exploring whether states' performance on this measure was related to their rates of inpatient hospitalization for mental health diagnoses (for the mental health denominator) or for alcohol and other drug use disorders (for the AOD denominator). We hypothesized that states' with lower rates of follow-up after discharge from the emergency department might have higher rates of inpatient stays for mental health and AOD. To evaluate the relationship between state performance on our measure and the state-level rate of inpatient stays, we fit a mixed effects logistic regression model. We regressed a beneficiary-level indicator of inpatient stay on a state-level binary variable indicating lowest vs. highest quartile performance follow-up after emergency department measure. To this we added a random effect of state to account for clustering of patients within states. If the p-value for the performance indicator variable is less than 0.05, then there is a significant difference in the rates of inpatient stays between states in the lowest vs. highest quartile of performance. If the p-value is greater than 0.05, then there is not a significant difference between low- and high-performing states.

Systematic Assessment of Face Validity

Our field test addressed the face validity of the measure specification by several types of stakeholder input. A multi-stakeholder technical expert panel of 16 individuals consisting of health plan representatives, behavioral health and quality measurement experts was convened and provided input throughout the measure development process, including review of the field test results and recommendations for final specifications.

In addition, four multi-stakeholder focus groups that included 29 representatives from Medicaid plans, states, integrated care systems, consumers/advocates, and other health care organizations reviewed and commented on the draft specifications and field test results.

We also received feedback from a two-week public comment period hosted on NCQA's online public comment system. The public comment notification was submitted to stakeholders representing consumers, health plans, clinicians, quality measurement and behavioral health experts.

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

2019 Submission

Results of face validity assessment

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

Statistical results of construct validity testing

Table 3a. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Emergency Department Visit for Mental Illness* Performance Scores Within Measure – **Commercial** Plans, 2017

	7-day follow-up
30-day follow-up	0.93

All scores were significant at p<0.05

Table 3b. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Emergency Department Visit for Mental Illness* Performance Scores Within Measure – **Medicaid** Plans, 2017

	7-day follow-up
30-day follow-up	0.96

All scores were significant at p<0.05

Table 3c. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* Performance Scores Within Measure – **Medicare** Plans, 2017

	7-day follow-up
30-day follow-up	0.92

All scores were significant at p<0.05

Table 4a. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Emergency Department Visit for Mental Illness* and *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* Measure Performance Scores – **Commercial** Plans, 2017

	Follow-Up After Emergency Department Visit for Mental Illness	
Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence	30 days	7 days
30-day follow-up	0.48	0.45
7-day follow-up	0.44	0.42

All scores were significant at p<0.05

Table 4b. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Emergency Department Visit for Mental Illness* and *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* Measure Performance Scores – **Medicaid** Plans, 2017

	Follow-Up After Emergency Department Visit for Mental Illness	
Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence	30 days	7 days
30-day follow-up	0.57	0.57
7-day follow-up	0.53	0.55

All scores were significant at p<0.05

Table 4c. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Emergency Department Visit for Mental Illness* and *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* Measure Performance Scores – **Medicare** Plans, 2017

	Follow-Up After Emergency Department Visit for Mental Illness	
Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence	30 days	7 days
30-day follow-up	0.48	0.49
7-day follow-up	0.42	0.44

All scores were significant at p<0.05

2014 Submission

Table 3: Utilization of Inpatient Hospitalization for Mental Health Diagnosis by Measure Performance Quartile

	Enrollees Hospitalized for Mental Health Diagnosis (Percentage)		
	Among States in Bottom 25 Percent of performance on FUED - Mental Health Denominator	Among States in Top 25 Percent of FUED – Mental Health Denominator	p-value
7-day follow-up	1.87%	1.79%	0.90
30-day follow-up	2.08%	1.72%	0.80

Table 4: Utilization of Inpatient Hospitalization for AOD Diagnosis by Measure Performance Quartile

	Enrollees Hospitalized for AOD Diagnosis (Percentage)		
	Among States in Bottom 25 Percent of FUED – AOD Denominator	Among States in Top 25 Percent of FUED - AOD Denominator	p-value
7-day follow-up	0.26%	0.32%	0.44
30-day follow-up	0.26%	0.32%	0.44

Systematic assessment of face validity
Focus group stakeholders and the technical expert panel both supported the face validity of the measure. Both groups agreed that the transition period post-emergency room discharge was a critical time to get patients into outpatient care. Of the stakeholders who provided public comment for this measure, 18 total comments were received and 13 (72.2%) supported or supported the measure with modifications. Other commenters who did not support the measure had concerns about identifying whether an emergency visit took place as well as the validity of the emergency department diagnosis. Specifically, stakeholders were concerned that if the diagnosis in formation is not received, follow-up There were additional concerns about the ability to act on the 7-day follow-up as there is lag time between the date of the visit and when the claim is received by the organization. However, our multi-stakeholder expert panel recommended moving forward with the measure because the specifications and testing results were reasonable and the measure addresses important quality opportunity.

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?*)

2019 Submission

Interpretation of systematic assessment of face validity
The multi-stakeholder advisory panels concluded the measures had good face validity.

Interpretation of construct validity testing
Correlations between individual rates within the *Follow-Up After Emergency Department Visit for Mental Illness* measure were strong (Tables 3a, 3b, 3c). Correlations between the *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* and *Follow-Up After Emergency Department Visit for Mental Illness* measure rates (Tables 4a, 4b, 4c) were moderate. Plans with higher rates on *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* tend to also have higher rates on the *Follow-Up After Emergency Department Visit for Mental Illness*. The results indicate that the *Follow-Up After Emergency Department Visit for Mental Illness* measure has good validity.

2014 Submission

While the empirical testing did not support our hypothesis, stakeholders generally supported the face validity of the measure. The rate of inpatient hospitalization is not statistically different between states that perform well on this measure versus states that perform poorly (Tables 3 and 4). However, this result is likely due to

the relatively low, tightly distributed rates of inpatient hospitalization for states in both the low- and high-performing groups. Rather than suggest that the measure is not valid, this result may indicate that our assumptions were not correct about the relationship between the measure and inpatient hospitalization; this relationship may warrant further study. The findings from public comment, focus groups and technical expert panel suggest that the adaptation for monitoring follow up after ED visits has specifications that can produce valid results.

_____

2b2. EXCLUSIONS ANALYSIS

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

2019 Submission

No exclusions

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*)

2014 Submission

Our testing addresses four components of the denominator or exclusions, as shown in Table 5.

Table 5: Measure Exclusions

Exclusion	Rationale	MH Denominator lost due to exclusion	AOD Denominator lost due to exclusion
ED discharges after December 1	If an ED discharge is after December 1, then the full 30-day follow-up period is not available for patient to receive follow-up care during the measurement year	7.5%	6.9%
ED discharges who die during the follow-up period	Death prevents follow-up care	Less than 1%	Less than 1%
For an ED discharge where the patient also visited the ED in the previous 30 days, exclude those previous ED discharges	Including these ED discharges could lead to a larger number of ED visits resulting in higher performance on the measure This exclusion aligns with the NQF-endorsed (#0576) Follow-up after Hospitalization for Mental Illness measure to reduce the burden and confusion for health plans implementing both measures	16.2%	17.3%
ED discharges with an inpatient or other residential stay during follow-up period	An inpatient or otherwise residential stay may interfere with the receipt of outpatient follow-up care This exclusion aligns with the NQF-endorsed (#0576) Follow-up after Hospitalization for Mental Illness measure to reduce the burden and confusion for health plans implementing both measures	34.2%	40.8%

Note: The exclusions presented in this table are not mutually exclusive. For example, a discharge that falls under exclusions 1 and 4 would appear in both places in this table.

We tested whether the exclusions affected over performance scores.

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*)

2014 Submission

Table 6: Number and percent of denominator remaining after exclusions, by state

	Mental Health (MH) Denominator			AOD Denominator		
State	MH denominator before exclusions	MH denominator after exclusions	Percent after exclusions	AOD denominator before exclusions	AOD denominator after exclusions	Percent after exclusions
AK	297	221	74.4%	294	212	72.1%
AL	3,244	2,294	70.7%	1,135	873	76.9%
CT	2,800	1,608	57.4%	2,081	1,135	54.5%
DC*	311	181	58.2%	302	0	0.0%
GA	5,009	3,506	70.0%	1,796	1,273	70.9%
IL	11,057	5,681	51.4%	3,179	1,248	39.3%
IN	1,405	990	70.5%	765	563	73.6%
KY	4,762	3,520	73.9%	1,879	1,403	74.7%
LA	3,738	2,447	65.5%	1,451	1,081	74.5%
MN	3,192	2,149	67.3%	1,100	747	67.9%
MS	1,198	842	70.3%	524	392	74.8%
NC	6,755	4,907	72.6%	3,372	2,416	71.6%
NH	800	574	71.8%	292	188	64.4%
OK	1,183	813	68.7%	717	514	71.7%
WI	1,491	1,041	69.8%	895	588	65.7%
WV	1,699	1,178	69.3%	934	704	75.4%
Total	48,941	31,952	65.3%	20,716	13,337	64.4%

*DC was dropped from AOD denominator due to small sample size.

Table 7: Measure performance before and after application of final exclusion

Measure	**Overall measure performance after exclusions 1-3 applied**	**Overall measure performance after exclusions 1-4 applied**
Mental Health: 7-day follow-up	66.6	67.8
Mental Health: 30-day follow-up	76.9	77.3
AOD: 7-day follow-up	64.2	66.6
AOD: 30-day follow up	67.9	68.7

Note: The overall performance rates presented here are pooled across states.

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*)

2014 Submission

We tested several exclusions in order to understand the impact on the denominator. Exclusions 1 and 2 are necessary to ensure that follow-up care can be observed during the measurement year. Exclusion 3 is prevents incentivizing more emergency department visits and aligns with other NQF endorsed measures to decrease burden and confusion for health plans. Average measure performance does not change substantially when Exclusion 4 is implemented, this exclusion aligns with NQF measure 0576, and there is a clinical rationale for excluding emergency department discharges that have an inpatient or other residential stay during the follow-up period, which is important to the face validity of the measure. All of the exclusions have minimal effect on the burden of calculating the measure since these exclusions are derived exclusively from claims data. In the specifications, some of these exclusions have been incorporated into the denominator definition.

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.
2019 Submission
N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure.

2b3.1. What method of controlling for differences in case mix is used?
☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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

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

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

2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:
☐ **Published literature**
☐ **Internal data analysis**
☐ **Other (please describe)**

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

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

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

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

2019 Submission

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

2014 Submission

Empirical testing

To demonstrate meaningful differences in performance, we calculated an inter-quartile range (IQR) for each rate. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, we calculate a Chi-squared test of the performance difference between each state in the lowest quartile vs. each state in the highest quartile. The Chi-squared test method calculates a test statistic based on the sample size and performance rate of each state. If the p value of the test statistic is less than .05, then the two states' performance is significantly different from each other. Using this method, we compared the performance rates of each pair of states, one state in the 25th percentile and another state in the 75th percentile of performance.

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

2019 Submission

Table 6. Calendar year 2017 Variation in Performance Across Health Plans

	Rate	Avg. EP	Avg. (%)	SD (%)	10th (%)	25th (%)	50th (%)	75th (%)	90th (%)	IQR (%)	p-value
Commercial	30-day follow-up	419	60.1	11.5	43.9	52.3	60.4	67.1	75.0	14.8	<0.05
	7-day follow-up	419	45.3	12.3	30.3	37.7	44.4	51.8	62.0	14.1	<0.05
Medicaid	30-day follow-up	1,114	54.8	14.7	37.8	45.6	52.8	66.3	74.5	20.7	<0.05
	7-day follow-up	1,114	40.1	15.4	21.9	28.9	37.3	53.0	61.3	24.1	<0.05
Medicare	30-day follow-up	166	47.6	13.5	31.3	37.8	47.0	56.6	65.0	18.8	<0.05
	7-day follow-up	166	31.5	13.6	16.3	21.7	29.0	39.1	49.6	17.4	<0.05

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

2014 Submission

Table 8: Variation in performance across states

Measure	10th	25th	Median	75th	90th	IQR	p-value
Mental Health: 7-day follow-up	46.0	67.0	74.8	80.7	89.4	13.7	<.001
Mental Health: 30-day follow-up	62.5	77.3	83.3	85.9	92.4	8.6	<.001
AOD: 7-day follow-up	32.8	61.1	72.1	82.4	90.3	21.4	<.001
AOD: 30-day follow up	34.1	62.6	74.8	82.5	90.3	19.9	<.001

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?*)

2019 Submission

The results above indicate there is a 14-24% gap in performance between the 25th and 75th performing plans. For all product lines and rates, the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for Medicaid plans 7-day follow-up rate, which show a 24.1 percentage point gap between 25th and 75th percentile plans. This gap represents an average of 268 more ED visits at each plan with follow-up in high performing plans compared to low performing plans (estimated from average health plan eligible encounters).

2014 Submission

The results above indicate there is a gap in performance between the 25th and 75th performing states, ranging from 8.6 percentage points on the 7-day mental health measure to 21.4 on the 7-day AOD measure. For all states and all rates, the difference between the 25th and 75th percentile is statistically significant.

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

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

2019 Submission

This measure has only one set of specifications.

Note: *This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to*

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

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

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

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

_____

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

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

2019 Submission

HEDIS measures apply to enrolled members in a health plan, and NCQA has a rigorous audit process to ensure the eligible population and numerator events for each measure are correctly identified and reported. The audit process is designed to verify primary data sources used to populate measures and ensure specifications are correctly implemented.

The HEDIS Compliance Audit addresses the following functions:

- Information practices and control procedures
- Sampling methods and procedures
- Data integrity
- Compliance with HEDIS specifications
- Analytic file production
- Reporting and documentation

2014 Submission

This measure is collected using all available administrative claims; there are no missing data on this measure.

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*)

2019 Submission

HEDIS addresses missing data in a structured way through its audit process. HEDIS measures apply to enrolled members in a health plan, and NCQA-certified auditors use standard audit methodologies to assess whether data sources are missing data. If a data source is found to be missing data, and the issues cannot be rectified, the auditor will assign a "materially biased" designation to the measure for that reporting plan, and the rate will not be used. Once measures are added to HEDIS, NCQA conducts a first-year analysis to assess the

measure's feasibility once widely implemented in the field. This analysis includes an assessment of how many plans report valid rates vs. rates that are materially biased. These considerations are weighed in the deliberation process before measures are approved for public reporting.

2014 Submission

Not applicable.

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*)

2019 Submission

This measure goes through the NCQA audit process each year to identify potential errors or bias in results. Only performances rates that have been reviewed and determined not to be "materially biased" are reported and used.

2014 Submission

Not applicable.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

3b.2. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources. For **maintenance of endorsement**, if this measure is not an eMeasure (eCQM), please describe any efforts to develop an eMeasure (eCQM).

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

Attachment:

3c. Data Collection Strategy

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

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

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

NCQA conducts an independent audit of all HEDIS collection and reporting processes, as well as an audit of the data which are manipulated by those processes, in order to verify that HEDIS specifications are met. NCQA has developed a precise, standardized methodology for verifying the integrity of HEDIS collection and calculation processes through a two-part program consisting of an overall information systems capabilities assessment followed by an evaluation of the MCO's ability to comply with HEDIS specifications. NCQA-certified auditors using standard audit methodologies will help enable purchasers to make more reliable comparisons between health plans.

The HEDIS Compliance Audit addresses the following functions:

1) Information practices and control procedures
2) Sampling methods and procedures
3) Data integrity
4) Compliance with HEDIS specifications
5) Analytic file production
6) Reporting and documentation

In addition to the HEDIS audit, NCQA provides a system to allow "real-time" feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system, NCQA responds immediately to questions and identifies possible errors or inconsistencies in the implementation of the measure. This system informs both annual updates to the measures as well as routine re-evaluation of measures. These processes include updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence.

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

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Regulatory and Accreditation Programs	Public Reporting https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html CMS Medicaid Adult Core Set NCQA Health Plan Ratings / Report Cards https://www.ncqa.org/hedis/reports-and-research/ratings-methodology-and-guidelines/ Quality Improvement (external benchmarking to organizations) NCQA Annual State of Health Care Quality http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality NCQA Quality Compass http://www.ncqa.org/hedis-quality-measurement/quality-measurement-products/quality-compass SAMHSA Demonstration Program for Certified Community Behavioral Health Clinics (CCBHCs) https://www.samhsa.gov/section-223

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

SAMHSA CERTIFIED COMMUNITY BEHAVIORAL HEALTH CLINICS: This is a demonstration program for states to certify community behavioral health clinics. Certified clinics must meet specific criteria emphasizing high-quality care including reporting quality measures.

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

HEALTH PLAN RATINGS/REPORT CARDS: This measure is used in the calculation of health plan ratings, which are reported on the NCQA website annually. These ratings are based on a plan's performance on their HEDIS, CAHPS and accreditation standards scores. In 2017, a total of 521 Medicare Advantage health plans, 614

commercial health plans and 294 Medicaid health plans across 50 states, D.C., Guam, Puerto Rico, and the Virgin Islands were included in the Ratings.
CMS MEDICAID ADULT CORE SET: There are a core set of health quality measures for Medicaid-enrolled adults. The Medicaid Adult Core Set was identified by the Centers of Medicare & Medicaid (CMS) in partnership with the Agency for Healthcare Research and Quality (AHRQ). The data collected from these measures will help CMS to better understand the quality of health care that adults enrolled in Medicaid receive nationally. Beginning in January 2014 and every three years thereafter, the Secretary is required to report to Congress on the quality of care received by adults enrolled in Medicaid. Additionally, beginning in September 2014, state data on the adult quality measures will become part of the Secretary's annual report on the quality of care for adults enrolled in Medicaid.
NCQA QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting health plans, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.
NCQA STATE OF HEALTH CARE ANNUAL REPORT: This measure is publicly reported nationally and by geographic regions in the NCQA State of Health Care annual report. This annual report published by NCQA summarizes findings on quality of care. In 2012, the report included measures on 11.5 Medicare Advantage beneficiaries in 455 Medicare Advantage health plans, 99.4 million members in 404 commercial health plans, and 14.3 million Medicaid beneficiaries in 136 plans across 50 states.

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A

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

N/A

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

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

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

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

NCQA publishes HEDIS results annually in our Quality Compass tool. NCQA also presents data at various conferences and webinars. For example, at the annual HEDIS Update and Best Practices Conference, NCQA presents results from all new measures' first year of implementation or analyses from measures that have changed significantly. NCQA also regularly provides technical assistance on measures through its Policy Clarification Support System, as described in Section 3c.1.

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

Describe how feedback was obtained.

NCQA measures are evaluated regularly using a consensus-based process to consider input from multiple stakeholders, including but not limited to entities being measured. We use several methods to obtain input, including vetting of the measure with several multi-stakeholder advisory panels, public comment posting, and review of questions submitted to the Policy Clarification Support System. This information enables NCQA to comprehensively assess a measure's adherence to the HEDIS Desirable Attributes of Relevance, Scientific Soundness and Feasibility.

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

Measure users have sought clarification on the types of encounters, as well as timing of encounters, that satisfy the measure. Measure users have sought clarification on the types of providers that qualify for the follow-up encounter. This feedback has helped us refine and clarify criteria in the measure specification.

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

This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in public reporting and quality improvement programs.

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.

During the measure's last major update, feedback obtained through the mechanisms described in 4a2.2.1 informed how we revised the measure to parse it out into two separate measures focused on follow up after and ED visit for mental health and alcohol use disorder, respectively.

Improvement

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

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

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

The performance rates were steady for Commercial and Medicaid plans and declined slightly for Medicare plans over the two years. This suggests the challenge in connecting members with mental illness to treatment after an ED visit. All health plans need to substantially improve follow-up care for mental health services. Enhancing connections between ED and outpatient services and increasing the mental health workforce are crucial components to improving access to follow-up care for this population.

4b2. Unintended Consequences

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

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

There were no identified unintended findings for this measure during testing or since implementation.

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

There were no identified unexpected benefits for this measure during testing or since implementation.

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)

0576 : Follow-Up After Hospitalization for Mental Illness (FUH)

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

5a. Harmonization of Related Measures

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

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

5a.2. If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.
The measure is harmonized with the existing NQF-endorsed measure. The following highlights the differences between the measures: Population focus (denominator): The measure targets patients discharged from the emergency department (not inpatient). Numerator: The measure captures follow-up with a primary mental health diagnosis (regardless of the type of provider).

5b. Competing Measures

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

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Not applicable.

Appendix

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

No appendix **Attachment:**

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): National Committee for Quality Assurance

Co.2 Point of Contact: Bob, Rehm, nqf@ncqa.org, 202-955-3500-

Co.3 Measure Developer if different from Measure Steward: National Committee for Quality Assurance

Co.4 Point of Contact: Kristen, Swift, Swift@ncqa.org, 202-955-5174-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

Behavioral Health Measurement Advisory Panel

Katharine Bradley, MD, MPH, Kaiser Permanente Washington Health Research Institute

Christopher Dennis, MD, MBA, FAPA, Landmark Health

Ben Druss MD, MPH, Emory University

Frank A. Ghinassi, PhD, ABPP, Rutgers University Behavioral Health Care

Constance M. Horgan, Sc.D., Brandeis University

Laura Jacobus-Kantor, PhD, SAMHSA, HHS

Jeffrey D. Meyerhoff, MD, Optum Behavioral Health

Harold Alan Pincus, MD, Irving Institute for Clinical and Translational Research --Columbia University

Michael Schoenbaum, PhD, National Institute of Mental Health

John H. Straus, MD, Beacon Health Options

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2014

Ad.3 Month and Year of most recent revision: 07, 2018

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

Ad.5 When is the next scheduled review/update for this measure? 12, 2020

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

Calculated measure results, based on unadjusted HEDIS specifications, may not be termed "Health Plan HEDIS rates" until they are audited and designated reportable by an NCQA-Certified Auditor. Such unaudited results should be referred to as "Unaudited Health Plan HEDIS Rates." Accordingly, "Heath Plan HEDIS rate" refers to

and assumes a result from an unadjusted HEDIS specification that has been audited by an NCQA-Certified HEDIS Auditor.

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

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

Ad.7 Disclaimers: This HEDIS® performance measure is not a clinical guideline and does not establish a standard of medical care and has not been tested for all potential applications.

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

Ad.8 Additional Information/Comments: NCQA Notice of Use. Broad public use and dissemination of these measures, without modification, are encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Modifications to, and/or commercial use of, a measure requires the prior written consent of NCQA and is subject to a license at the discretion of NCQA. As used herein, “commercial use” refers to any sale, license, or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed, or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.