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

Purple text represents the responses from measure developers.

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

Brief Measure Information

NQF #: 0576

Corresponding Measures:

De.2. Measure Title: Follow-Up After Hospitalization for Mental Illness

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

De.3. Brief Description of Measure: The percentage of discharges for members 6 years of age and older who were hospitalized for treatment of selected mental illness or intentional self-harm diagnoses and who had a follow-up visit with a mental health provider. Two rates are reported:

1. The percentage of discharges for which the member received follow-up within 30 days after discharge.
2. The percentage of discharges for which the member received follow-up within 7 days after discharge.

1b.1. Developer Rationale: This measure assesses whether health plan members who were hospitalized for a mental illness or intentional self-harm received a timely follow-up visit. Follow-up care following an acute event, such as hospitalization, reduces the risk of negative outcomes (e.g., medication errors, re-admission, emergency department use). Efforts to facilitate treatment following a hospital discharge also lead to less attrition in the initial post-acute period of treatment. Thus, this time period may be an important opportunity for health plans to implement strategies aimed at establishing strong relationships between patients and mental health providers and facilitate ongoing engagement in treatment.

According to an analysis of data from the National Inpatient Sample (NIS), between 2007 and 2014 there were over 1.5 million nonfatal suicide attempts requiring hospitalization, a rate of 67.1 per 100,000 persons (Connor et al., 2019). Another analysis of the NIS found that of 122,574 hospital discharges in 2003 with an injury diagnosis, 7.6% were for intentional self-harm (Patrick et al., 2010).

Fontanella et al. (2020) examined the association between timely outpatient follow-up after a psychiatric hospitalization and risk of death by suicide, and found that youths with a follow-up visit within 7 days of discharge had a significantly lower risk of death by suicide. A study of 90-day readmissions among individuals with schizophrenia and bipolar disorder found that individuals with an outpatient visit within 30-days following discharge experienced a lower risk of readmission within the following 90 days (Marcus et al., 2017). Similarly, Mark and colleagues (2013) found that increased follow-up at community mental health centers was associated with lower risk of re-admission among Medicaid patients hospitalized for mental illness or substance use disorder.

Evidence suggests that brief, low-intensity interventions are effective in bridging the gap between inpatient and outpatient treatment (Dixon 2009) and improving patient experience of continuity of care (Tomita &

Herman, 2015). Low-intensity interventions are typically implemented at periods of high risk for treatment dropout, such as following an emergency room or hospital discharge or the time of entry into outpatient treatment. For example, Boyer et al evaluated strategies aimed at increasing attendance at outpatient appointments following hospital discharge. They found that the most common factor in a patient's medical history that was linked to a patient having a follow-up visit was a discussion about the discharge plan between the inpatient staff and outpatient clinicians. Other strategies they found that increased attendance at appointments included having the patient meet with outpatient staff and visit the outpatient program prior to discharge (Boyer 2000).

Barekatain M, Maracy MR, Rajabi F, Baratian H. (2014). Aftercare services for patients with severe mental disorder: A randomized controlled trial. J Res Med Sci. 19(3):240-5.

Boyer, C. A., McAlpine, D. D., Pottick, K. J., & Olfson, M. (2000). Identifying risk factors and key strategies in linkage to outpatient psychiatric care. The American journal of psychiatry, 157(10), 1592–1598. https://doi.org/10.1176/appi.ajp.157.10.1592

Conner, A., Azrael, D., & Miller, M. (2019). Suicide Case-Fatality Rates in the United States, 2007 to 2014: A Nationwide Population-Based Study. Annals of internal medicine, 171(12), 885–895. https://doi.org/10.7326/M19-1324

Dixon L, Goldberg R, Iannone V, et al. Use of a critical time intervention to promote continuity of care after psychiatric inpatient hospitalization for severe mental illness. Psychiatr Serv. 2009;60:451–458.

Fontanella, C. A., Warner, L. A., Steelesmith, D. L., Brock, G., Bridge, J. A., & Campo, J. V. (2020). Association of Timely Outpatient Mental Health Services for Youths After Psychiatric Hospitalization With Risk of Death by Suicide. JAMA network open, 3(8), e2012887.

Kreyenbuhl, J., Nossel, I., & Dixon, L. (2009). Disengagement from mental health treatment among individuals with schizophrenia and strategies for facilitating connections to care: A review of the literature. Schizophrenia Bulletin, 35, 696-703.

Luxton DD, June JD, Comtois KA. (2013). Can postdischarge follow-up contacts prevent suicide and suicidal behavior? A review of the evidence. Crisis. 34(1):32-41. doi: 10.1027/0227-5910/a000158.

Marcus, S. C., Chuang, C. C., Ng-Mak, D. S., & Olfson, M. (2017). Outpatient Follow-Up Care and Risk of Hospital Readmission in Schizophrenia and Bipolar Disorder. Psychiatric services (Washington, D.C.), 68(12), 1239–1246.

Mark, T., Tomic, K. S., Kowlessar, N., Chu, B. C., Vandivort-Warren, R., & Smith, S. (2013). Hospital Readmission Among Medicaid Patients with an Index Hospitalization for Mental and/or Substance Use Disorder. The Journal of Behavioral Health Services & Research, 40(2), 207–221.

Patrick, A. R., Miller, M., Barber, C. W., Wang, P. S., Canning, C. F., & Schneeweiss, S. (2010). Identification of hospitalizations for intentional self-harm when E-codes are incompletely recorded. Pharmacoepidemiology and drug safety, 19(12), 1263–1275. https://doi.org/10.1002/pds.2037

Tomita, A., & Herman, D. B. (2015). The role of a critical time intervention on the experience of continuity of care among persons with severe mental illness after hospital discharge. The Journal of nervous and mental disease, 203(1), 65–70.

S.4. Numerator Statement: 30-Day Follow-Up: A follow-up visit with a mental health provider within 30 days after discharge.

7-Day Follow-Up: A follow-up visit with a mental health provider within 7 days after discharge.

S.6. Denominator Statement: Discharges from an acute inpatient setting with a principal diagnosis of mental illness or intentional self-harm on the discharge claim during the first 11 months of the measurement year (i.e. January 1 to December 1) for members 6 years and older.

S.8. Denominator Exclusions: Exclude from the denominator for both rates, members who begin using hospice services anytime during the measurement year (Hospice Value Set)

Exclude both the initial discharge and the readmission/direct transfer discharge if the readmission/direct transfer discharge occurs after December 1 of the measurement year.

Exclude discharges followed by readmission or direct transfer to a nonacute facility within the 30-day follow-up period regardless of principal diagnosis.

Exclude discharges followed by readmission or direct transfer to an acute facility within the 30-day follow-up period if the principal diagnosis was not for mental health or intentional self-harm.

These discharges are excluded from the measure because rehospitalization or transfer may prevent an outpatient follow-up visit from taking place.

De.1. Measure Type: Process

S.17. Data Source: Claims

S.20. Level of Analysis: Health Plan

IF Endorsement Maintenance – Original Endorsement Date: Dec 04, 2009 **Most Recent Endorsement Date:** Jun 28, 2017

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

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

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

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

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

The developer provides the following evidence for this measure:

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

Summary of prior review in 2016

- This is a maintenance process measure using claims data at the plan level that assesses the percentage of discharges for members 6 years of age and older who were hospitalized for treatment of selected mental illness or intentional self-harm diagnoses and who had a follow-up visit with a mental health provider. Two rates are reported: follow-up within 30 days and within 7 days after discharge.
- In the 2016 submission, the developer cited the following guidelines that support follow-up after hospitalization:
    - National Institute for Health and Care Excellence (NICE) Guideline – Schizophrenia (published 2009)
        - This guideline was not graded; no evidence offered to support the recommendation.
    - NICE – Psychosis and Schizophrenia (published 2014)
        - The developer suggests that the guideline was graded using the GRADE approach, however the submission is unclear if or how the follow-up recommendation was graded. No evidence offered to support the recommendation.
    - American Psychiatric Association (APA) Guideline – Schizophrenia (published 2004)
        - The guideline was graded as [I] recommended with substantial clinical confidence and [II] recommended with moderate clinical confidence
    - APA Guidelines – Bipolar Disorder (published 2002)
        - The guideline was graded as [I] recommended with substantial clinical confidence
    - APA Guidelines – Major Depressive Disorder (published 2010)
        - The guideline was graded as [I] recommended with substantial clinical confidence

Changes to evidence from last review

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

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

Updates:

- In the current submission, the developer provided updates to the following guidelines provided in the 2016 submission:
    - APA Guidelines – Schizophrenia (updated in 2019)
        - The guideline recommends combining pharmacotherapy and psychosocial interventions to treat patients with a possible psychotic disorder or those with schizophrenia
        - It is unclear how the guideline and recommendation were graded
- The developer added the NICE Guideline on the transition between inpatient mental health settings and community or care home settings
    - Published in 2016, the guideline recommends discussing follow-up with the person before discharge and to follow-up with a person who has been discharged within 7 days
    - The evidence was graded as Moderate (+) or Poor to Moderate (-/+) evidence and Moderate (+) to Good (++) evidence, where:
        - ++ indicates all or most of the checklist criteria have been fulfilled, and where they have not been fulfilled, the conclusions are unlikely to alter
        - + indicates that some of the checklist criteria have been fulfilled, and where they have not been fulfilled or adequately described, the conclusions are unlikely to alter
        - – indicates that few or no checklist criteria have been fulfilled and the conclusions are likely or very likely to alter
    - 10 studies were included in the evidence statements, which is indicative of a high quantity
        - Of the studies, 6 were randomized control trials (RCTs) and 4 were qualitative studies; 5 studies were rated moderate, 4 were rated good, and 1 was rated poor quality.
- All evidence consistently shows follow-up care reduces suicide attempts, readmissions, and improves functioning

- Developer proffers new evidence within the submission that cites Marcus et al, 2017, a study which found that for patients with schizophrenia and bipolar disorder, outpatient follow up within 30 days post-discharge was associated with lower readmission risk. "Outpatient visits during the 30 days after discharge were associated with a lower hospital readmission risk during the following 90 days. Assertive hospital discharge planning to secure outpatient visits after hospital discharge is needed for these patient populations."

Questions for the Committee:

- The evidence provided by the developer is updated and directionally the same compared to that for the previous NQF review. Does the Committee agree?
- The evidence provided does include the recommended follow-up within 30 days of discharge. Are you aware of other evidence to support the 30-day rate?

Guidance from the Evidence Algorithm

Box 1 the measure does not assess performance on a health outcome Box 3 evidence matches what is being measured Box 4 QQC provided Box 5b MODERATE

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

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

Maintenance measures – increased emphasis on gap and variation

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

- For commercial health plans, the developer presented the following mean performance rates:
  - In 2018, the mean 7-day rate was 0.44 and the mean 30-day rate was 0.6
  - In 2017, the mean 7-day rate was 0.46 and the mean 30-day rate was 0.68
- For Medicare health plans, the developer presented the following mean performance rates:
  - In 2018, the mean 7-day rate was 0.28 and the mean 30-day rate was 0.48
  - In 2017, the mean 7-day rate was 0.32 and the mean 30-day rate was 0.53
- For Medicaid health plans, the developer presented the following mean performance rates:
  - In 2018, the mean 7-day rate was 0.36 and the mean 30-day rate was 0.57
  - In 2017, the mean 7-day rate was 0.37 and the mean 30-day rate was 0.58
- There is high variability across the different product lines, demonstrating room for improvement among health plans

Disparities

- Disparity data was not provided

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Are you aware of evidence that disparities exist in this area of healthcare?

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

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

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures–are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- The new evidence cites the APA guideline recommending combining pharmacotherapy and psychosocial interventions to treat patients with a possible psychotic disorder or those with schizophrenia. This measure is only focused on the follow-up visit not the pharmacotherapy. In addition it's not clear whether the follow-up mental health visits included an EBP therapeutic practice. The evidence provided does include the recommended follow-up within 30 days of discharge.
- Evidence for impact of substance use disorders on intentional self-harm is not included. Which demonstrates significant evidence gap that threatens validity. See work of RRies, et al, University of Washington.
- Evidence is sufficient. Would be interested in more evidence to support 30-day follow up.
- The evidence presented by the developer is directly related to the process measure. I am not aware of new studies that are not cited in this submission.
- Follow up after hospitalization has been a part evidence-based practice for most major mental illnesses for decades.
- Evidence is directionally consistent. However, there is much less evidence about 7 days versus 30. Also, the components necessary to count as effective follow up are not well delineated. I didn't see the inclusion of telehealth, primary care (particularly relevant in more rural areas), or home visit interventions.
- Moderate, addition of Marcus 2017 paper further strengthens the evidence
- Evidence applies directly to process being measured. Outpatient follow-up visits within 30 days is associated with lower readmission risk. I am not aware of any new information that changes the evidence base for this measure that has not been cited in the submission.
- Empirical data cited by submission applies directly to the measure (Marcus 2017, updated APA guidelines). Follow-up care has been shown to reduce rate of re-hospitalization & costs. Vidal et al (2020) showed use of ACT following inpatient discharge decreased rehospitalization and Smith et al (2019) found that contact between inpatient discharge planners and outpatient clinics improved rates of 7-day and 30-day aftercare.
- The updated evidence solidifies the previous evidence that timely follow-up DOES make a difference. Psych Services Dec 2-17 Outpatient Follow-Up Care and Risk of Hospital Readmission in Schizophrenia and Bipolar Disorder shows reduced readmissions with follow up in 30 days.
- This is a process measure determining the impact of follow-up visits within 7 days or within 30 day post discharge for psychiatric patients ages 6 years and older. The evidence they have provided is directly applicable to the purpose of the outcome measure. They believe that visits within 7 days or within 30 days can reduce improve engagement with care, reduce readmissions and risk of self-harm, and they provide evidence that demonstrates this. am not aware of any other studies or information that changes the evidence base for this measure.
- Agree that the evidence provided by the developer is updated to support and directionally the same compared to that for the previous NQF review. Appreciate additional 30-day evidence and not aware of other evidence to support 30-day rate.

- Process

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- Performance Gap is High across different payers. A lot of room for improvement. Data on disparities is not provided.
- The absence of disparity data is concerning.
- Sufficiently demonstrated a performance gap. Unless, I missed it, I did not see data on population subgroups, so I was not able to evaluate disparities in care by subpopulations.
- Yes. And from the data presented the developer clearly demonstrates a gap in care and therefore opportunities for improvement. Disparity data was not provided and the reason cited was that social risk factor data were not available, but is actively engaged to integrate social risk.
- The submission noted that there is high variability across the different product lines, demonstrating room for improvement among health plans.
- There is a gap. The fact that disparities data are not included is astonishing and a clear deficit. Side note: we should encourage, perhaps require disparities data as a must pass criterion.
- Yes
- There is high variability across the different health plan product lines. Disparity data was not provided.
- There is a high gap in care, as the mean 7-day follow-up rate was 19% and the 30-day rate was 29%. Data on gender and race were provided by the measure developer and indicate gaps in care that could be further studied.
- There is a significant performance gap. It also looks like the measure results worsened slightly between 2017 and 2018 along ALL product lines . Can we ask the developer to comment on this please? Slightly
- Yes, performance data was provided comparing rates of 7-day follow up and rates of 30-day follow up for commercial payers, Medicare and Medicaid. There is clearly a difference in services provided to patients based on payer type, a performance gap that warrants a national performance measure. Data on disparities was not provided.
- Does seem to be sizable variability across product lines, suggesting room for improvement. No disparities data provided – and not personally aware of evidence around disparity data in this are - but would welcome the chance to learn more if such evidence becomes available.
- Yes data provided for all lines of business. Subgroups by line of business yes by demographics no. Information on disparities not provided.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

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

Reliability

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

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

Validity

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

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

Composite measures only:

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

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

Evaluators: NQF BHSU Staff

Summary of NQF Staff Evaluation

Reliability

- The developer used a Beta-binomial model to measure the signal-to-noise ratio. The signal is defined as the proportion of variability in measured performance that can be explained by real differences across the reporting entities (in this case, health plans). The developer estimated reliability for each reporting entity and then averaged the reliability estimates across all reporting entities to produce a point estimate of signal-to-noise reliability (labeled "mean signal-to-noise reliability"). It measures how well, on average, the measure can differentiate between reporting entity performance on the measure. The Beta-Binomial methodology is a common approach used by measure developers to establish the confidence that a given provider sample has been ranked appropriately. This is considered a standard approach.
- The developer provided standard error (SE) and 95% confidence interval (95% CI) of the mean signal-to-noise reliability and stratified by the denominator size.
- The developer provided the distribution of the plan-level signal-to-noise reliability estimates.
- The developer provided point estimates of mean signal-to-noise reliability by health plan type. Full results can be found in section 2a2.3. The results suggest the measure has high reliability.
    - Table 2a of section 2a2.3 provides the point estimate of mean signal-to-noise reliability, its SE, and the 95% CI for the 7-day measure rate for each type of health plan, stratified by denominator size.

        - The reliability estimate for commercial plans is 0.884 with a 95% CI of (0.872, 0.895). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.
        - The reliability estimate for Medicaid plans is 0.969 with a 95% CI of (0.962, 0.975). Stratified analyses show that reliability exceeds 0.9 for all terciles.
        - The reliability estimate for Medicare plans 0.900 with a 95% CI of (0.890, 0.909). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.
    - Table 2b of section 2a2.3 provides the point estimate of mean signal-to-noise reliability, its SE, and the 95% CI for the 30-day measure rate for each type of health plan, stratified by denominator size.
        - The reliability estimate for commercial plans is 0.883 with a 95% CI of (0.872, 0.895). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.
        - The reliability estimate for Medicaid plans is 0.967 with a 95% CI of (0.960, 0.975). Stratified analyses show that reliability exceeds 0.9 for all terciles.
        - The reliability estimate for Medicare plans is 0.910 with a 95% CI of (0.902, 0.918). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.

Validity

- The developer assessed both construct (convergent) validity and face validity.
- For construct validity:
    - Validity for this measure was measured against NCQA's *Follow-Up After Emergency Department Visit for Mental Illness (FUM)* since both measures address similar populations and activities for patients following an acute event involving mental illness.
    - The developer performed Pearson correlation analysis for construct validity, which estimates the strength of linear association between two continuous variables.
        - P-values were calculated to determine the significance of the correlation coefficients, with the threshold set at 0.05, where values less than this imply it is unlikely that a non-zero coefficient was observed due to chance alone, i.e. it is significant.
- The developer reported Pearson correlation results within the measure between 7-day and 30-day. The correlations were positive and strong across product lines.
    - For Commercial plans, the correlation coefficient is 0.90 with significance at $p < 0.001$ (table 4a of section 2b1.3).
    - For Medicaid plans, the correlation coefficient is 0.93 with significance at $p < 0.001$ (table 4b of section 2b1.3).
    - For Medicare plans, the correlation coefficient is 0.91 with significance at $p < 0.001$ (table 4c of section 2b1.3).
- The developer also reported Pearson correlation results between *Follow-Up After Hospitalization for Mental Illness (FUH)* and *Follow-Up After Emergency Department Visit for Mental Illness (FUM)* (table 5a of section 2b1.3). The correlations were positive and moderate.

- Face validity was conducted and was systematically determined by the developer's Committee on Performance Measurement. The multi-stakeholder advisory panels concluded the measures had good face validity.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

The Committee did not recommend NQF 3572 Follow-Up After Psychiatric Hospitalization (FAPH) (Centers for Medicare and Medicaid Services/Mathematica). BHSU Committee's validity concern for NQF 3572 and have rated the validity as moderate due to this concern. Note that the developer has pointed out that only the last admission should be counted when a readmission occurs within 30 days.

- Developer has provided no exclusion data or analysis.
- The Committee should discuss the implications of the exclusion related to this measure and assess whether the same rationale should or should not be applied to NQF 0576, and what it means that the last discharge should be considered.

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

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

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- added telehealth. Curiously, it excludes visits on date of discharge. Also the definition of provider does not include FQHCs or other primary care providers who increasingly have integrated BH expertise. Also, licensed clinical alcohol and drug counselors are not included in the definition of provider.
- Yes appears reliable/repeatable with high CI.
- No concerns.
- Data elements are clearly defined. The 30 day discharge measure is reasonable while I do have concerns about the likelihood the 7-day as is demonstrated in the data presented.
- The testing sample included a broad range of commercial, Medicaid, and Medicare plans, so no concerns about reliability.
- SNR--moderate to high. Wish there was risk/case adjustment analysis
- Usual NCQA approach because they limit testing reliability based only on their data source

- Data elements are clearly defined, with adequate sampling, and good reliability demonstrated through a standard approach. I do not have concerns about the likelihood that this measure can be consistently implemented.
- I have no concerns about implementation as the data used to measure are claims data.
- Measure basically looks reliable. Literature suggests that 7 day follow-up for adolescents is lower for blacks/other minorities, older adolescents, those with medical co-morbidities, and for those with diagnosis other than Depression, Schizophrenia, or Bipolar Affective Disorder.
- Data elements are clearly defined. Codes with descriptors are provided and the steps are clear. I do not have concerns that this measure can be consistently implemented as reliability estimates are high.
- Measure is generally well-defined to ensure consistency here. I remain unclear what providers would "count" as a follow-up visit. Specifically, would this include a visit with a primary care provider who is not a mental health provider but who is helping manage the mental health condition (e.g. via medication management)?
- Reliability is high no need to discuss.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- Highly reliable
- I am concerned about the lack of disparity data.
- No.
- No
- No
- No, it's ok
- above
- No because signal-to-noise, SE, CI for each type of health plan were all were within acceptable range.
- No
- No
- I do not.
- Beta-Binomial methodology to measure signal-to-noise ratio seems adequate, though this is new to me. I don't see a need to discuss reliability testing, but also defer to those with more experience here.
- Reliability is high no need to discuss.

2b1. Validity -Testing: Do you have any concems with the testing results?

- Moderate
- Comparing this to FUM data interesting--so why do we need two measures for the same thing?
- No.
- No. But I did not quite understand the Committee's previous validity concern for NQF3572.
- No
- I hate our approach to validity testing--the idea that two measures have high correlations may simply reflect that neither is valid. Certainly, the fact there is moderate correlation to mental health fu and general ER fu is not unexpected--but it says little about whether they are truly valid. OK, I understand this is a common psychometric approach to validity—enough said.
- Limited to construct validity comparing to another conceptually similar measure developed by NCQA. Face validity based on I assume a survey item completed by the advisory panel the developer convened during development of measure

- I have concerns about patients who are excluded from the measure.
- No
- Nothing significant
- I do not.
- Pearson correlation analysis for construct seems adequate. I don't see a need to discuss validity testing, but also defer to those with more experience here.
- No concerns.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- no concerns
- I think this measure should be bifurcated. First part: FU after MH: SCPT, BAD, MDD is valid. Second part: FU after intentional self-harm is not substantiated based upon the Evidence 1a above or should have exclusions.
- Acceptable.
- This is incomplete because the developers do not present the data by social risk factors.
- The exclusions are consistent with the evidence. Risk adjustment is appropriately developed and tested.
- Exclusions. Again, I wish there was some scenario analysis of different SDOH.
- no, but this is likely related to NCQA limiting to their data source
- Per my answer in Q7, I am concerned about patients excluded from the measure, specifically those being transferred to another facility and the concern that the rehospitalization may interfere with their outpatient follow up. Telemedicine and/or good coordination of care could obviate this potential barrier. Risk adjustment - n/a.
- Exclusions are consistent and appropriate.
- The decision to exclude anyone readmitted in the first 30 days post discharge allows for cleaner measurement but also eliminates the most severe individuals. This does not invalidate the measure but probably makes the intervention look somewhat more effective than it truly is.
- HEDIS does not report exclusions, but there are exclusions based on readmissions. Need more clarity and discussion around the exclusions. Does not appear that specific groups are excluded inappropriately. HEDIS does not include information on risk adjustment as it is not applicable.
- No concerns. HEDIS Compliance Audit addresses missing data. IQR check addresses meaningful differences. Comparability N/A as measure has only one set of specifications. Risk adjustment or stratification N/A.
- No risk adjustment in this measure.

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about

quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- No concerns
- I have concerns with construct validity because this measure does not address the impact of substance use/abuse on intentional self-harm; nor is there evidence provided in the section 1a. Valid when substance misuse was the primary cause.
- I am satisfied with the testing of the measure's validity.
- I do not have concerns regarding these possible threats to validity.
- No
- We should discuss exclusions, briefly
- If you apply this measure to Medicaid claims data, you may underestimate follow-up care for children and youth if mental health care was provided in schools or funded by state funds earmarked for certain types of publicly-funded outpatient mental health care (e.g. assessment for special ed services)
- IQR and t-test demonstrated plans' performance is significantly different. Comparability - n/a. Missing data does not appear to be a threat to validity.
- No
- Nothing significant.
- HEDIS includes data on follow-up visits across commercial plans, Medicare and Medicaid. This measure identified meaningful differences in performance between health plans. This data is very useful. There are not multiple sets of specifications, and the data go through an audit process to address any missing data. "Materially biased" information is excluded.
- No. HEDIS Compliance Audit addresses this adequately.
- No threat to validity.

Criterion 3. Feasibility

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

3. **Feasibility** is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
   - The measure data is coded by someone other than person obtaining original information
   - All of the data elements are in defined fields in a combination of electronic sources
   - Noncommercial use of the measure does not require consent by the measure developer
   - Commercial use of the measure requires written consent from the developer

Questions for the Committee:

- Do you have concerns about the feasibility of this measure?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. **Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?**
   - The measure is feasible-uses claims data
   - EHR data - no concerns.
   - It's claims data, so it is feasible. Any concern that delays in data entry or claims disputes could impact results? Do we have information that supports that most claims from community-based mental health providers are submitted electronically?
   - I do not have concerns regarding this as for this measure health systems will be able to measure the required data elements.
   - This is a limited process measure and easy to operationalize.
   - Feasible.
   - highly feasible
   - No concerns regarding feasibility.
   - No concerns since this measure uses claims data and has been used for over 10 years
   - It is feasible
   - Some data are generated and used during care delivery, but others would be collected after the care episode. This requires multiple systems for data collection, all of which use electronic forms. I understand its current use is for quality improvement and is being used by many groups and in various projects, but the process of implementation/use is not clear to me.
   - No concerns.
   - No concerns this is straight forward.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

OR

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

Accountability program details

- This measure is currently used in several CMS programs including:
  - Medicaid Child Core Set
  - Medicaid Adult Core Set
  - Hospital Compare
  - Qualified Health Plan (QHP) Quality Rating System (QRS)
  - Quality Payment Program (QPP)
- The measure is also used for NCQA's accreditation of commercial, Medicaid, and Medicare plans, Health Plan Ratings/Report Cards, State of Health Care Annual Report
- Other programs where the measure is in use include: Quality Compass and Inpatient Psychiatric Facility Quality Reporting

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

Feedback on the measure by those being measured or others

- The developer conducts reevaluation regularly
- Though the developer has received clarifying questions on specifications (e.g. whether a certain type of provider met the definition of mental health providers), they report that health plans have considered the measure feasible for reporting

Additional Feedback:

- Feedback received has informed the developer's revisions to the measure specifications to include clarifying text and additional examples to further support determining numerator compliance

Questions for the Committee:

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

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

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

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

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

Improvement results

- 2017 to 2018 data demonstrates relatively stable improvement but also shows that there is room for improvement across health plans
- The large variation between 10th and 90th percentiles also suggest room for improvement across health plans

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

Unexpected findings (positive or negative) during implementation

- None found

Potential harms

- **None**

Additional Feedback:

***Questions for the Committee**:*

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

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

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

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- yes.
- n/a
- Yes
- I have no concerns about accountability or transparency. The feedback mechanism for those being measured is described.
- The measure is publicly reported by CMS. The submission notes that there is substantial feedback.
- used widely; fine.
- existing measure, one of the easier ones for state Medicaid agencies to report
- Measure is being publicly reported and is used in several CMS programs, NCQA's accreditation process, and in psychiatric programs as a quality tool. Those being measured have been provided with the results and have been invited to participate in feedback/measure implementation.
- Data have been collected and reported publicly for 10 years; feedback is routinely collected and used to clarify measure definitions.
- It's currently being used and I am unaware of problems with it.
- Health plans submit their results to NCQA. NCQA then publicly reports rates (publish, present at conferences and webinars) across all plans and benchmarks to help plans compare their performance. Yes, the results are disclosed and available. Accountability applications are in 4.1 and 4a1.1 (health plan ratings, quality of care, accreditation, and more). Entities that use the HEDIS are aware of their performance results and data prior to submitting their data to NCQA. Since data is publicly reported by NCQA, entities are able to compare their performance to others through benchmarking. It is not clear what assistance is provided beyond comparing their results to others and technical assistance through

NCQA's Clarification Support System. There are opportunities to provide feedback on the measures through advisory panels, public comment posting and review of questions submitted. This feedback was used in recent revisions, which included clarifying text and providing additional examples.

- My only question is the extent to which health plans are explicitly encouraged/incentivized to share data and coordinate with delivery system entities to advance improvement. Do we have a sense that this is already happening?
- Yes

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- The benefits would be increased if the measure included information about whether the follow-up encounter included pharmacotherapy and evidence based therapeutic.
- I have concern about this as a public performance measure because of lack of disparity data. Maybe useful as a quality improvement measure.
- Measure is usable.
- I believe that the performance results must have healthcare systems think about models of care that are patient-centered for high-risk individuals. This will hopefully reduce barriers for care for a vulnerable population and I believe the benefits would outweigh any theoretical harms.
- The benefits of follow-up after hospitalization far outweigh any potential harm.
- ok
- none
- 2017 to 2018 data demonstrates relatively stable improvement which help further the goal of efficient healthcare by increasing patient tenure in the community and facilitating ongoing management of their illness through outpatient providers. I believe the potential benefits outweigh the harms in this measure.
- N/A. This is an important issue and results still show opportunity for improvement that outweigh any potential harm.
- Would like the developer to explain small worsening of results between first and second year's results.
- If follow-up visits within a specific time frame do improve engagement in care, reduce risks of self-harm and reduce readmissions as the evidence demonstrates, then this tool can be used to further high-quality efficient healthcare. Comparing follow-ups by payer and/or site of service and engaging in quality improvement and accreditation processes, to name a few, the quality of care provided can be improved and services utilized more efficiently to meet the desired goals. This will help payers determine coverage and sites which services to provide based on data. Since the measure relies on claims data for services already rendered, there are no unintended consequences that I can identify. Rather, this measure will provide the potential benefits of improving care provided and efficiently use resources, assuming that entities use the data to implement meaningful change.
- Given limited access to and availability of outpatient BH care, am curious whether there has been any examination of whether systems seeking to improve FUM have ever done so at the expense of FUH (or vice versa).
- None

Criterion 5: Related and Competing Measures

Related or competing measures
None
Harmonization
N/A

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

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- No information provided.
- Yes. doesn't FUM measure compete?
- none
- N/A
- There are no competing measures.
- N/A
- no
- n/a
- N/A
- Nothing per se. Could we ask developer if they could note any commonalities in interventions that seem to work for both Follow up after ED Visit and this measure??
- There are no related or competing measures.
- Do not appear to be any related/competed measures. Same question as above - given limited access to and availability of outpatient BH care, am curious whether there has been any examination of whether systems seeking to improve FUM have ever done so at the expense of FUH (or vice versa).
- None

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/15/2021

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

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 0576

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

Type of measure:

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

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

Data Source:

☒ **Claims** ☐ **Electronic Health Data** ☐ **Electronic Health Records** ☐ **Management Data**
☐ **Assessment Data** ☐ **Paper Medical Records** ☐ **Instrument-Based Data** ☐ **Registry Data**
☐ **Enrollment Data** ☐ **Other**

Level of Analysis:

☐ **Clinician: Group/Practice** ☐ **Clinician: Individual** ☐ **Facility** ☒ **Health Plan**
☐ **Population: Community, County or City** ☐ **Population: Regional and State**
☐ **Integrated Delivery System** ☐ **Other**

Measure is:

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

RELIABILITY: SPECIFICATIONS

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

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

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

   - The measure score reliability was calculated from Healthcare Effectiveness Data and Information Set (HEDIS) data that included 358 commercial health plans, 172 Medicaid plans, and 308 Medicare plans.
   - The testing sample included all commercial, Medicaid, and Medicare plans submitting data to NCQA for this HEDIS measure.

2. **Briefly summarize any concerns about the measure specifications.**
   - None identified by staff.

RELIABILITY: TESTING

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

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

   ☐ **Yes** ☐ **No**

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

   Submission document: Testing attachment, section 2a2.2

   - The developer used a Beta-binomial model to measure the signal-to-noise ratio. The signal is defined as the proportion of variability in measured performance that can be explained by real differences across the reporting entities (in this case, health plans).
     - The score ranges from 0.0 to 1.0, with 0.0 implying that all variation is attributed to measurement error (i.e. noise), and 1.0 implying that all variation is caused by real differences in performance across health plans.

        - A score of 0 indicates none of the variation (signal) is attributable to the plan.
        - A score of 0.7 or higher is often suggested to indicate adequate reliability to distinguish performance between two plans.
        - A score of 1.0 indicates all of the variation (signal) is attributable to the plan.
    - The developer estimated reliability for each reporting entity and then averaged the reliability estimates across all reporting entities to produce a point estimate of signal-to-noise reliability (labeled "mean signal-to-noise reliability"). It measures how well, on average, the measure can differentiate between reporting entity performance on the measure.
- The developer provided standard error (SE) and 95% confidence interval (95% CI) of the mean signal-to-noise reliability and stratified by the denominator size.
- The developer provided the distribution of the plan-level signal-to-noise reliability estimates.
- The Beta-Binomial methodology is a common approach used by measure developers to establish the confidence that a given provider sample has been ranked appropriately. This is considered a standard approach.

7. **Assess the results of reliability testing**
    - The developer provided point estimates of mean signal-to-noise reliability by health plan type. Full results can be found in section 2a2.3. The results suggest the measure has high reliability.
        - Table 2a of section 2a2.3 provides the point estimate of mean signal-to-noise reliability, its SE, and the 95% CI for the 7-day measure rate for each type of health plan, stratified by denominator size.
            - The reliability estimate for commercial plans is 0.884 with a 95% CI of (0.872, 0.895). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.
            - The reliability estimate for Medicaid plans is 0.969 with a 95% CI of (0.962, 0.975). Stratified analyses show that reliability exceeds 0.9 for all terciles.
            - The reliability estimate for Medicare plans 0.900 with a 95% CI of (0.890, 0.909). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.
        - Table 2b of section 2a2.3 provides the point estimate of mean signal-to-noise reliability, its SE, and the 95% CI for the 30-day measure rate for each type of health plan, stratified by denominator size.
            - The reliability estimate for commercial plans is 0.883 with a 95% CI of (0.872, 0.895). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.
            - The reliability estimate for Medicaid plans is 0.967 with a 95% CI of (0.960, 0.975). Stratified analyses show that reliability exceeds 0.9 for all terciles.
            - The reliability estimate for Medicare plans is 0.910 with a 95% CI of (0.902, 0.918). The stratified analyses indicate that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.

    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.**

    Box 1 → Measure specifications are unambiguous and complete → Box 2 → Empirical reliability testing was conducted using statistical tests → Box 4 → Reliability testing was conducted with computed performance measure scores → Box 5 → Signal-to-noise reliability estimates were calculated, as well as SE, 95% CI, and distribution of the signal-to-noise reliability estimates → Box 6a → All reliability estimates exceeded the 0.7 threshold for reliability → HIGH

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Testing attachment, section 2b2.

    - The developer does not collect data on exclusions for HEDIS reporting of the measure.
    - The measure specification has some exclusions that the BHSU Committee has found concerning in other follow-up measures:
        - Exclude discharges followed by readmission or direct transfer to a nonacute facility within the 30-day follow-up period regardless of principal diagnosis.
        - Exclude discharges followed by readmission or direct transfer to an acute facility within the 30-day follow-up period if the principal diagnosis was not for mental health or intentional self-harm.
    - Developer notes that these discharges are excluded from the measure because rehospitalization or transfer to a nonacute facility may prevent an outpatient follow-up visit from taking place. If the readmission/direct transfer to an acute facility was for a principal diagnosis of mental health or intentional self-harm, count only the last discharge.
    - The developer excludes patients whose outcome (readmission) the measure focus is designed to address. The Committee did not recommend NQF 3572 Follow-Up After Psychiatric Hospitalization (FAPH) (Centers for Medicare and Medicaid Services/Mathematica).

- There were concerns that exclusions may impact the validity of the measure since the measure excludes those who have undesirable outcomes that could be due to lack of follow-up, which represented a significant portion of the target population (35%).
- This measure also includes opioid use disorder. Committee noted that follow-up is to prevent readmission and death, especially for opiate use disorder.
- Only the last discharge is counted.

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

    Submission document: Testing attachment, section 2b4.

    - No concerns; developer appropriately tested for meaningful differences.
        - The developer calculated an inter-quartile range (IQR) for each indicator, which provides a measure of the dispersion of performance. It is the difference between the 25th and 75th percentile on a measure. An independent sample t-test was performance to assess the statistical significance of the difference.
        - The developer reported the percentage point gaps for each health plan type and provided it in section 2b4.2. The difference in performance between plans is statistically significant across all health plan entities.

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.

    - No concerns; the measure has a single data source and only one set of specifications.

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

    Submission document: Testing attachment, section 2b6.

    - HEDIS measures go through an audit process to ensure data for each measure are correctly identified and reported.

16. **Risk Adjustment**

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

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

    ☐ Yes ☐ No ☒ Not applicable

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

    - The developer assessed both construct (convergent) validity and face validity.
    - For construct validity:
        - Validity for this measure was measured against NCQA's *Follow-Up After Emergency Department Visit for Mental Illness (FUM)* since both measures address similar populations and activities for patients following an acute event involving mental illness.

- The developer performed Pearson correlation analysis for construct validity, which estimates the strength of linear association between two continuous variables.
  - Values range from -1 to +1
    - A value of 1 indicates a strong positive linear association, i.e. an increase in one variable is associated with an increase of another variable.
    - A value of 0 indicates no linear association.
    - A value of -1 indicates a strong negative linear association, i.e. an increase in one variable is associated with a decrease of another variable.
  - P-values were calculated to determine the significance of the correlation coefficients, with the threshold set at 0.05, where values less than this imply it is unlikely that a non-zero coefficient was observed due to chance alone, i.e. it is significant.
- Face validity was conducted and was systematically determined by the developer's Committee on Performance Measurement.

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

    Submission document: Testing attachment, section 2b2.3

    - The developer reported Pearson correlation results within the measure between 7-day and 30-day. The correlations were positive and strong across product lines.
      - For Commercial plans, the correlation coefficient is 0.90 with significance at $p < 0.001$ (table 4a of section 2b1.3).
      - For Medicaid plans, the correlation coefficient is 0.93 with significance at $p < 0.001$ (table 4b of section 2b1.3).
      - For Medicare plans, the correlation coefficient is 0.91 with significance at $p < 0.001$ (table 4c of section 2b1.3).
    - The developer also reported Pearson correlation results between *Follow-Up After Hospitalization for Mental Illness (FUH)* and *Follow-Up After Emergency Department Visit for Mental Illness (FUM)* (table 5a of section 2b1.3). The correlations were positive and moderate.
    - The multi-stakeholder advisory panels concluded the measures had good face validity.

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

    Submission document: Testing attachment, section 2b1.

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (score-level testing was not performed)

22. **Was the method described and appropriate for assessing the accuracy of ALL critical data elements?**

    NOTE that data element validation from the literature is acceptable.

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

    ☐ **Yes**

    ☐ **No**

    ☒ **Not applicable** (data element testing was not performed)

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

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

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

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

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

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

    Box 1 → all potential threats to validity were empirically tested → Box 2 → Pearson correlation analyses were conducted → Box 5 → Yes → Box 6 → Correlation of the performance measure scores were compared to a similar performance measure → Box 7a → HIGH

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.**

    Staff are concerned about the follow-up exclusion in light of the BHSU Committee's validity concern for NQF 3572 and have rated the validity as moderate due to this concern.

    - Developer has provided no exclusion data or analysis.
    - The Committee should discuss the implications of the exclusion related to this measure and assess whether the same rationale should or should not be applied to NQF 0576, and what it means that the last discharge should be considered.

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

xxxxxxxxxx.docx

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

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

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0576

Measure Title: Follow-Up After Hospitalization for Mental Illness

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

Date of Submission: 11/2/2020

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

Outcome

☐ Outcome:

☐ Patient-reported outcome (PRO):

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

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

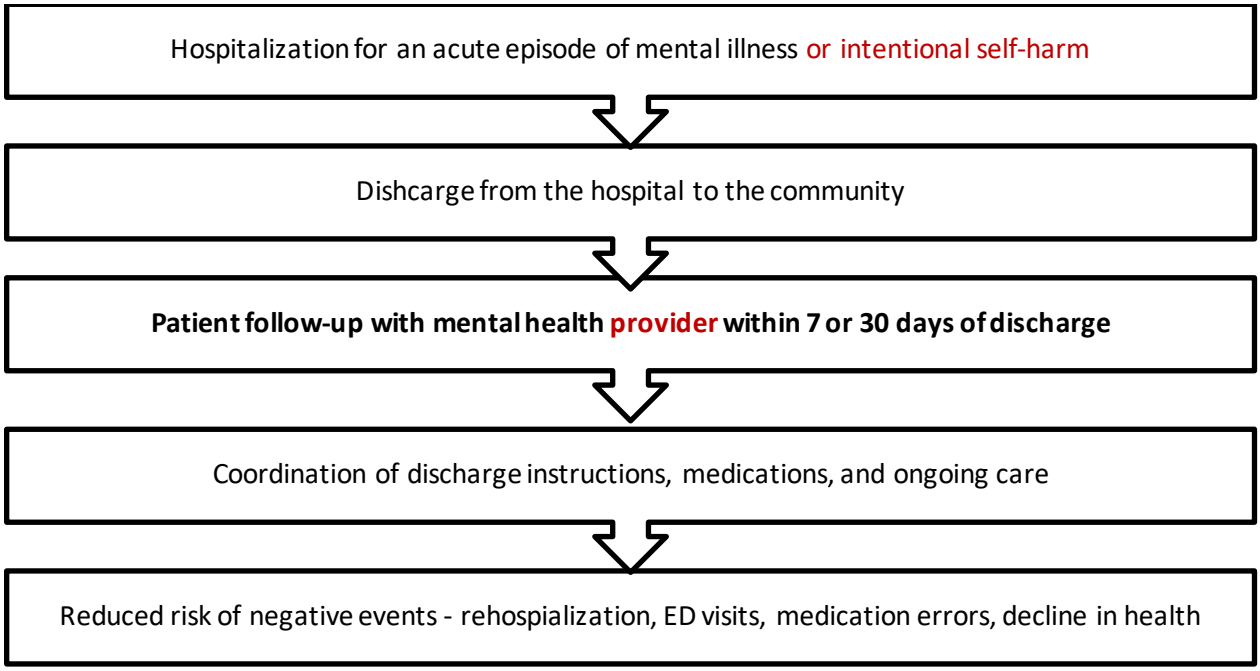
☒ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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



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

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

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

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

National Institute for Health and Care Excellence (NICE) Guideline - Transition between inpatient mental health settings and community or care home settings

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	Transition between inpatient mental health settings and community or care home settings 2016 National Institute for Health and Care Excellence. Transition between inpatient mental health settings and community or care home settings. . London (UK): National Institute for Health and Care Excellence (NICE); 2016 Aug. 22 p. (NICE clinical guideline; NG53). https://www.nice.org.uk/guidance/ng53/resources/transition-between-inpatient-mental-health-settings-and-community-or-care-home-settings-pdf-1837511615941
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	1.6.1. Discuss follow-up support with the person before discharge. Arrange support according to their mental and physical health needs. This could include: contact details, for example of: • a community psychiatric nurse or social worker • the out-of-hours service • support and plans for the first week • practical help if needed • employment support. 1.6.7. Follow up with a person who has been discharged within 7 days.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2020 submission:** 1.6.1. Moderate (+) or Poor to Moderate (-/+)evidence 1.6.7. Moderate (+) to Good (++) evidence

Systematic Review	Evidence
	++ All or most of the checklist criteria have been fulfilled, and where they have not been fulfilled the conclusions are very unlikely to alter. + Some of the checklist criteria have been fulfilled, and where they have not been fulfilled, or are not adequately described, the conclusions are unlikely to alter. – Few or no checklist criteria have been fulfilled and the conclusions are likely or very likely to alter.
Provide all other grades and definitions from the evidence grading system	**2020 submission:** ++ All or most of the checklist criteria have been fulfilled, and where they have not been fulfilled the conclusions are very unlikely to alter. + Some of the checklist criteria have been fulfilled, and where they have not been fulfilled, or are not adequately described, the conclusions are unlikely to alter. – Few or no checklist criteria have been fulfilled and the conclusions are likely or very likely to alter.
Grade assigned to the **recommendation** with definition of the grade	**2020 submission:** Recommendation statements not graded.
Provide all other grades and definitions from the recommendation grading system	**2020 submission:** Recommendation statements not graded
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2020 submission:** 10 studies were included in the evidence statements corresponding to these recommendation statements related to follow-up support. 6 were RCTs and 4 were qualitative studies. 5 were rated moderate, 4 were rated good, and 1 was rated poor.
Estimates of benefit and consistency across studies	**2020 submission:** "The absence of relevant, high quality recent effectiveness studies in arriving at these principles of care meant that it was not possible to ascertain and compare trade-off between benefits and harms for people in implementing these recommendations."
What harms were identified?	**2020 submission:** "The absence of relevant, high quality recent effectiveness studies in arriving at these principles of care meant that it was not possible to ascertain and compare trade-off between

Systematic Review	Evidence
	benefits and harms for people in implementing these recommendations."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2020 submission:** N/A

National Institute for Health and Care Excellence (NICE) Guideline - Schizophrenia

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	2020 submission: No updates. This guideline has not been updated. 2016 Submission: Schizophrenia: core interventions in the treatment and management of schizophrenia in adults in primary and secondary care National Collaborating Centre for Mental Health 2009 National Collaborating Centre for Mental Health. Schizophrenia: core interventions in the treatment and management of schizophrenia in adults in primary and secondary care. London (UK): National Institute for Health and Clinical Excellence (NICE); 2009 Mar. 41 p. (NICE clinical guideline; no. 82). http://guidelines.gov/content.aspx?id=14313
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** Getting Help Early • Healthcare professionals should facilitate access as soon as possible to assessment and treatment, and promote early access throughout all phases of care. Initiation of Treatment (First Episode) Early Referral

Systematic Review	Evidence
	• Urgently refer all people with first presentation of psychotic symptoms in primary care to a local community-based secondary mental health service (for example, crisis resolution and home treatment team, early intervention service, community mental health team). Referral to early intervention services may be from primary or secondary care. The choice of team should be determined by the stage and severity of illness and the local context. • Carry out a full assessment of people with psychotic symptoms in secondary care, including an assessment by a psychiatrist. Write a care plan in collaboration with the service user as soon as possible. Send a copy to the primary healthcare professional who made the referral and the service user. • Include a crisis plan in the care plan, based on a full risk assessment. The crisis plan should define the role of primary and secondary care and identify the key clinical contacts in the event of an emergency or impending crisis. Early Post-Acute Period In the early period of recovery following an acute episode, service users and healthcare professionals will need to jointly reflect upon the acute episode and its impact, and make plans for future care.
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** Guideline was not graded.
Provide all other grades and definitions from the evidence grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A
Grade assigned to the **recommendation** with definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A

Systematic Review	Evidence
Provide all other grades and definitions from the recommendation grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A
Estimates of benefit and consistency across studies	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A
What harms were identified?	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission** N/A

National Institute for Health and Care Excellence (NICE) Guideline – Psychosis and Schizophrenia

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	2020 submission: No updates. NICE "checked this guideline in March 2019 and found no new evidence that affects the recommendations in this guideline." Thus, the guideline was not updated. 2016 Submission:

Systematic Review	Evidence
	Psychosis and schizophrenia in adults: treatment and management. 2014 National Collaborating Centre for Mental Health. Psychosis and schizophrenia in adults: prevention and management. London (UK): National Institute for Health and Care Excellence (NICE); 2014 Mar. 58 p. (NICE clinical guideline; no 178). https://www.nice.org.uk/guidance/cg178/resources/psychosis-and-schizophrenia-in-adults-prevention-and-management-35109758952133
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2020 submission:** No updates. This guideline has not been updated. **2016 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** • 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]**

Systematic Review	Evidence
	1.3.3 Assessment and care planning • 1.3.3.1 Carry out a comprehensive multidisciplinary assessment of people with psychotic symptoms in secondary care. This should include assessment by a psychiatrist, a psychologist or a professional with expertise in the psychological treatment of people with psychosis or schizophrenia. 1.4.6 Early post-acute period • 1.4.6.1 After each acute episode, encourage people with psychosis or schizophrenia to write an account of their illness in their notes. **[2009]** • 1.4.6.2 Healthcare professionals may consider using psychoanalytic and psychodynamic principles to help them understand the experiences of people with psychosis or schizophrenia and their interpersonal relationships. **[2009]** • 1.4.6.3 Inform the service user that there is a high risk of relapse if they stop medication in the next 1–2 years. **[2009]** • 1.4.6.4 If withdrawing antipsychotic medication, undertake gradually and monitor regularly for signs and symptoms of relapse. **[2009]** 1.4.6.5 After withdrawal from antipsychotic medication, continue monitoring for signs and symptoms of relapse for at least 2 years. **[2009]**
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 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 organization 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 summarize 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

Systematic Review	Evidence
	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 categorized into one of four groups (high, moderate, low, very low). https://www.nice.org.uk/guidance/cg178/evidence/appendix-13-490503567
Provide all other grades and definitions from the evidence grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 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 organization 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 summarize both the quality of the evidence and the results of the evidence

Systematic Review	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 categorized into one of four groups (high, moderate, low, very low). https://www.nice.org.uk/guidance/cg178/evidence/appendix-13-490503567
Grade assigned to the **recommendation** with definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 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.

Systematic Review	Evidence
	In these instances, give a reference to supporting documents. These recommendations apply to all patients. • recommendations for interventions that should (or should not) be used: For recommendations on interventions that 'should' be used, the GDG is confident that, for the vast majority of people, the intervention (or interventions) will do more good than harm, and will be cost effective. • recommendations for interventions that could be used: For recommendations on interventions that 'could' be used, the GDG is confident that the intervention will do more good than harm for most patients, and will be cost effective Recommendations are marked as **[2009]**, **[2009, amended 2014]**, **[2014]** or **[new 2014]**. • **[2009]** indicates that the evidence has not been reviewed since 2009. • **[2009, amended 2014]** indicates that the evidence has not been reviewed since 2009 but changes have been made to the recommendation wording that change the meaning. • **[2014]** indicates that the evidence has been reviewed but no changes have been made to the recommendation. **[new 2014]** indicates that the evidence has been reviewed and the recommendation has been updated or added.
Provide all other grades and definitions from the recommendation grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 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.

Systematic Review	Evidence
	In these instances, give a reference to supporting documents. These recommendations apply to all patients. • recommendations for interventions that should (or should not) be used: For recommendations on interventions that 'should' be used, the GDG is confident that, for the vast majority of people, the intervention (or interventions) will do more good than harm, and will be cost effective. • recommendations for interventions that could be used: For recommendations on interventions that 'could' be used, the GDG is confident that the intervention will do more good than harm for most patients, and will be cost effective Recommendations are marked as **[2009]**, **[2009, amended 2014]**, **[2014]** or **[new 2014]**. • **[2009]** indicates that the evidence has not been reviewed since 2009. • **[2009, amended 2014]** indicates that the evidence has not been reviewed since 2009 but changes have been made to the recommendation wording that change the meaning. • **[2014]** indicates that the evidence has been reviewed but no changes have been made to the recommendation. • **[new 2014]** indicates that the evidence has been reviewed and the recommendation has been updated or added.
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** NICE guideline recommendations are based on the best available evidence. We use a wide range of different types of evidence and other information – from scientific research using a variety of methods, to testimony from practitioners and people using services.
Estimates of benefit and consistency across studies	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:**

Systematic Review	Evidence
	All primary-level studies included after the first scan of citations were acquired in full and re-evaluated for eligibility at the time they were being entered into the study information database. More specific eligibility criteria were developed for each review question and are described in the relevant clinical evidence chapters. Eligible systematic reviews and primary-level studies were critically appraised for methodological quality (risk of bias) using a checklist (see The Guidelines Manual (NICE, 2012b) for templates). The eligibility of each study was confirmed by at least one member of the GDG.
What harms were identified?	No identified harms are cited.
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2020 submission:** No updates. **2016 Submission:** Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

American Psychiatric Association (APA) Guideline- Schizophrenia

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	2020 submission: The American Psychiatric Association Practice Guideline For The Treatment Of Patients With Schizophrenia Third Edition American Psychiatric Association 2019 American Psychiatric Association (2019). Practice Guideline for the Treatment of Patients With Schizophrenia Third Edition; 2019 Dec. 184 p. https://psychiatryonline.org/doi/full/10.1176/appi.books.9780890424841.Schizophrenia03 2016 Submission: Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition American Psychiatric Association 2004

Systematic Review	Evidence
	American Psychiatric Association (2004). Practice Guideline for the Treatment of Patients With Schizophrenia Second Edition; 2004 Feb. 184 p. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/schizophrenia.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2020 submission:** **Assessment and Determination of Treatment Plan** 1. APA *recommends* (1C) that the initial assessment of a patient with a possible psychotic disorder include the reason the individual is presenting for evaluation; the patient's goals and preferences for treatment; a review of psychiatric symptoms and trauma history; an assessment of tobacco use and other substance use; a psychiatric treatment history; an assessment of physical health; an assessment of psychosocial and cultural factors; a mental status examination, including cognitive assessment; and an assessment of risk of suicide and aggressive behaviors, as outlined in APA's *Practice Guidelines for the Psychiatric Evaluation of Adults* (3rd edition). 2. APA *recommends* (1C) that patients with schizophrenia have a documented, comprehensive, and person-centered treatment plan that includes evidence-based nonpharmacological and pharmacological treatments. **Pharmacotherapy** 1. APA *recommends* (1A) that patients with schizophrenia be treated with an antipsychotic medication and monitored for effectiveness and side effects.* *This guideline statement should be implemented in the context of a person-centered treatment plan that includes evidence-based nonpharmacological and pharmacological treatments for schizophrenia. **Psychosocial Interventions** 1. APA *recommends* (1B) that patients with schizophrenia who are experiencing a first episode of psychosis be treated in a coordinated specialty care program.* 2. APA *recommends* (1B) that patients with schizophrenia be treated with cognitive-behavioral therapy for psychosis (CBTp).* 3. APA *recommends* (1B) that patients with schizophrenia receive supported employment services.*

Systematic Review	Evidence
	4. APA *recommends* (1B) that patients with schizophrenia receive assertive community treatment if there is a history of poor engagement with services leading to frequent relapse or social disruption (e.g., homelessness; legal difficulties, including imprisonment).* 5. APA *suggests* (2C) that patients with schizophrenia receive interventions aimed at developing self-management skills and enhancing person-oriented recovery.* 6. APA *suggests* (2C) that patients with schizophrenia who have a therapeutic goal of enhanced social functioning receive social skills training.* **2016 Submission:** Stable Phase [A, A-, B, C, D, E, F, G] "Treatment programs need to combine medications with a range of psychosocial services to reduce the need for crisis-oriented hospitalizations and emergency department visits and enable greater recovery [I]." Acute Phase Treatment [A, A-, B, C, D, E, F, G] "It is recommended that pharmacological treatment be initiated promptly, provided it will not interfere with diagnostic assessment, because acute psychotic exacerbations are associated with emotional distress, disruption to the patient's life, and a substantial risk of dangerous behaviors to self, others, or property [I]." Acute Phase Treatment [A, A-, B, C, D, E, F, G] "Psychosocial interventions in the acute phase are aimed at reducing overstimulating or stressful relationships, environments, or life events and at promoting relaxation or reduced arousal through simple, clear, coherent communications and expectations; a structured and predictable environment; low performance requirements; and tolerant, nondemanding, supportive relationships with the psychiatrist and other members of the treatment team. Providing information to the patient and the family on the nature and management of the illness that is appropriate to the patient's capacity to assimilate information is recommended [II]. Patients can be encouraged to collaborate with the psychiatrist in selecting and adjusting the medication and other treatments provided [II]."

Systematic Review	Evidence
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2020 submission:** High (denoted by the letter A) = high confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect. Moderate (denoted by the letter B) = moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate. Low (denoted by the letter C) = low confidence that the evidence reflects the true effect. Further research is likely to change our confidence in the estimate of effect and is likely to change the estimate. **2016 Submission:** The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline. Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.

Systematic Review	Evidence
	[C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Provide all other grades and definitions from the evidence grading system	**2020 submission:** No other grades. **2016 Submission:** The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline (see Section VI). Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.

Systematic Review	Evidence
	[C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Grade assigned to the **recommendation** with definition of the grade	**2020 submission:** Each guideline statement is separately rated to indicate strength of recommendation and strength of supporting research evidence. Strength of recommendation describes the level of confidence that potential benefits of an intervention outweigh potential harms. This level of confidence is a consensus judgment of the authors of the guideline and is informed by available evidence, which includes evidence from clinical trials as well as expert opinion and patient values and preferences. There are two possible ratings: recommendation or suggestion. A **recommendation** (denoted by the numeral 1 after the guideline statement) indicates confidence that the benefits of the intervention clearly outweigh harms. A **suggestion** (denoted by the numeral 2 after the guideline statement) indicates greater uncertainty. **2016 Submission:** [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence.
Provide all other grades and definitions from the recommendation grading system	**2020 submission:** Strength of recommendation describes the level of confidence that potential benefits of an intervention outweigh potential harms. This level of confidence is a consensus judgment of the authors of the guideline and is informed by available evidence,

Systematic Review	Evidence
	which includes evidence from clinical trials as well as expert opinion and patient values and preferences. There are two possible ratings: recommendation or suggestion. A recommendation (denoted by the numeral 1 after the guideline statement) indicates confidence that the benefits of the intervention clearly outweigh harms. A suggestion (denoted by the numeral 2 after the guideline statement) indicates greater uncertainty. **2016 Submission:** Each recommendation is identified as falling into one of three categories of endorsement, indicated by a bracketed Roman numeral following the statement. The three categories represent varying levels of clinical confidence regarding the recommendation: [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence. [III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2020 submission:** "The Agency for Healthcare Research and Quality's (AHRQ) systematic review *Treatments for Schizophrenia in Adults* (McDonagh et al. 2017) served as the predominant source of information for this guideline. Databases that were searched are Ovid MEDLINE® (PubMed®), the Cochrane Central Register of Controlled Trials, the Cochrane Database of Systematic Reviews, and PsycINFO®. Results were limited to English-language, adult (18 and older), and human-only studies." "Recent, comprehensive, good- or fair-quality systematic reviews served as a primary source of evidence, supplemented by information from randomized controlled trials (RCTs) published since the systematic reviews or when no systematic reviews were available. For assessment of harms of treatment, systematic reviews of observational trials were also included. Eligibility for inclusion and exclusion of articles adhered to preestablished criteria. Specifically, the AHRQ review included articles that had at least 12 weeks of follow-up and were conducted in outpatient settings in countries that were relevant to the United States' health care system." "For key question 1 on antipsychotic treatment, 698 citations were identified, 519 of which were excluded on the basis of title and abstract review, yielding 179 full-text articles that were reviewed, of which **38 were included in the final AHRQ**

Systematic Review	Evidence
	review. For key question 2 on psychosocial and other nonpharmacological interventions, 2,766 citations were identified, 1,871 of which were excluded on the basis of title and abstract review, yielding 895 full-text articles that were reviewed, of which **53 were included in the final AHRQ review**." **2016 Submission:** "Relevant literature was identified through a computerized search of PubMed for the period from 1994 to 2002. Using the keywords schizophrenia OR schizoaffective, a total of 20,009 citations were found. After limiting these references to clinical trials and meta-analyses published in English that included abstracts, 1,272 articles were screened by using title and abstract information. The Cochrane Database of Systematic Reviews was also searched by using the keyword schizophrenia. Additional, less formal literature searches were conducted by APA staff and individual members of the work group on schizophrenia. Sources of funding were considered when the work group reviewed the literature but are not identified in this document. When reading source articles referenced in this guideline, readers are advised to consider the sources of funding for the studies"
Estimates of benefit and consistency across studies	**2020 submission:** "A *recommendation* (denoted by the numeral 1 after the guideline statement) indicates confidence that the benefits of the intervention clearly outweigh harms. A *suggestion* (denoted by the numeral 2 after the guideline statement) indicates greater uncertainty. Although the benefits of the statement are still viewed as outweighing the harms, the balance of benefits and harms is more difficult to judge, or the benefits or the harms may be less clear. When a negative statement is made, ratings of strength of recommendation should be understood as meaning the inverse of the above (e.g., *recommendation* indicates confidence that harms clearly outweigh benefits)." **Assessment and Determination of Treatment Plan** **Benefits** "In an individual with a possible psychotic disorder, a detailed assessment is important in establishing a diagnosis, recognizing co-occurring conditions (including substance use disorders,

Systematic Review	Evidence
	other psychiatric disorders, and other physical health disorders), identifying psychosocial issues, and developing a plan of treatment that can reduce associated symptoms, morbidity, and mortality." "Development and documentation of a comprehensive, person-centered treatment plan assures that the clinician has considered the available nonpharmacological and pharmacological options for treatment and has identified those treatments that are best suited to the needs of the individual patient, with a goal of improving overall outcome. It may also assist in forming a therapeutic relationship, eliciting patient preferences, permitting education about possible treatments, setting expectations for treatment, and establishing a framework for shared decision-making. Documentation of a treatment plan promotes accurate communication among all those caring for the patient and can serve as a reminder of prior discussions about treatment." "The potential benefits of this guideline statement were viewed as far outweighing the potential harms." **Pharmacotherapy** **Benefits** "Use of an antipsychotic medication in the treatment of schizophrenia can improve positive and negative symptoms of psychosis (high strength of research evidence) and can also lead to reductions in depression and improvements in quality of life and functioning (moderate strength of research evidence). A meta-analysis of double-blind, randomized, placebo-controlled trials showed a medium effect size for overall efficacy (Leucht et al. 2017), with the greatest effect on positive symptoms. The rates of achieving any response or a good response were also significantly greater in patients who received an antipsychotic medication. In addition, the proportion of individuals who dropped out of treatment for any reason and for lack of efficacy was significantly less in those who were treated with an antipsychotic medication. Research evidence from head-to-head comparison studies and network meta-analysis (McDonagh et al. 2017) showed no consistent evidence that favored a specific antipsychotic medication, with the possible exception of clozapine." "The potential benefits of this guideline statement were viewed as far outweighing the potential harms."

Systematic Review	Evidence
	Psychosocial Interventions **Benefits** Across all forms of psychosocial interventions recommended or suggested in this guideline, the APA concludes that potential benefits are likely to outweigh potential harms. Benefits cited include, reduced likelihood of relapse, reduced core illness symptoms, reduced symptom severity, and improved quality of life. **2016 Submission:** "The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline"
What harms were identified?	**2020 submission:** **Assessment and Determination of Treatment Plan** **Harms** "Harms may include serious adverse events; less serious adverse events that affect tolerability; minor adverse events; negative effects of the intervention on quality of life; barriers and inconveniences associated with treatment; and other negative aspects of the treatment that may influence decision-making by the patient, the clinician, or both. Some individuals may become anxious, suspicious, or annoyed if asked multiple questions during the evaluation. This could interfere with the therapeutic relationship between the patient and the clinician." **Pharmacotherapy** **Harms** "The harms of using an antipsychotic medication in the treatment of schizophrenia include sedation, side effects mediated through dopamine receptor blockade (e.g., acute dystonia, akathisia, parkinsonism, tardive syndromes, NMS,

Systematic Review	Evidence
	hyperprolactinemia), disturbances in sexual function, anticholinergic effects, weight gain, glucose abnormalities, hyperlipidemia, orthostatic hypotension, tachycardia, and QTc prolongation. Clozapine has additional harms associated with its use, including sialorrhea, seizures, neutropenia (which can be severe and life-threatening), myocarditis, and cardiomyopathy. Among the antipsychotic medications, there is variability in the rates at which each of these effects occurs, and no specific medication appears to be devoid of possible side effects." **Psychosocial Interventions** **Harms** Across all psychosocial interventions recommended, the APA concludes that the potential harms are not well documented but are likely to be minimal. **2016 Submission:** "The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2020 submission:** We are not aware of any further systematic reviews or studies published since the publication of this guideline that contraindicate the need for appropriate follow-up after hospitalization for mental illness. **2016 Submission:** Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

American Psychiatric Association (APA) Guidelines-Bipolar Disorder

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	2020 submission: No updates. This guideline has not been updated. 2016 Submission: Practice Guideline for the Treatment of Patients With Bipolar Disorder, Second Edition American Psychiatric Association 2002 American Psychiatric Association (2002) Practice Guideline for the Treatment of Patients With Bipolar Disorder, Second Edition; 2002 Apr. 82 p. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/bipolar.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** Psychiatric Management [A, C, D, E, F, G] "Specific goals of psychiatric management include establishing and maintaining a therapeutic alliance, monitoring the patient's psychiatric status, providing education regarding bipolar disorder, enhancing treatment compliance, promoting regular patterns of activity and of sleep, anticipating stressors, identifying new episodes early, and minimizing functional impairments [I]."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2020 submission:** No updates. **2016 Submission:** The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline (see Section VI). Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that

Systematic Review	Evidence
	readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments. [C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Provide all other grades and definitions from the evidence grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline (see Section VI). Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments.

Systematic Review	Evidence
	[A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial. [C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Grade assigned to the **recommendation** with definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** [I] Recommended with substantial clinical confidence.
Provide all other grades and definitions from the recommendation grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** Each recommendation is identified as falling into one of three categories of endorsement, indicated by a bracketed Roman numeral following the statement. The three categories represent varying levels of clinical confidence regarding the recommendation: [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence. [III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies?	**2020 submission:** No updates. This guideline has not been updated.

Systematic Review	Evidence
• Quality – what type of studies?	**2016 Submission:** "A computerized search of the relevant literature from MEDLINE and PsycINFO was conducted. Sources of funding were not considered when reviewing the literature. The first literature search was conducted by searching MEDLINE and PsycINFO for the period from 1992 to 2000. Key words used were "bipolar disorder," "bipolar depression," "mania," "mixed states," etc. A total of 122 citations were found. A search on PubMed was also conducted through 2001 that used the search terms "electroconvulsive," "intravenous drug abuse," "treatment response," "pharmacogenetic," "attention deficit disorder," "violence," "aggression," "aggressive," "suicidal," "cognitive impairment," "sleep," "postpartum," "ethnic," "racial," "metabolism," "hyperparathyroidism," "overdose," "toxicity," "intoxication," "pregnancy," "breast-feeding," and "lactation." Additional, less formal, literature searches were conducted by APA staff and individual members of the work group on bipolar disorder"
Estimates of benefit and consistency across studies	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** "The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline."
What harms were identified?	"The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline."
Identify any new studies conducted since the SR. Do the new studies	Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this

Systematic Review	Evidence
change the conclusions from the SR?	guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

American Psychiatric Association (APA) Guidelines-Major Depressive Disorder

Systematic Review	Evidence
Source of Systematic Review: • Title • Author • Date • Citation, including page number • URL	2020 submission: No updates. This guideline has not been updated. 2016 Submission: Practice Guideline for the Treatment of Patients With Major Depressive Disorder, Third Edition American Psychiatric Association 2010 American Psychiatric Association (2010); 2004 Practice Guideline for the Treatment of Patients With Major Depressive Disorder, Third Edition. 2010 Oct. 151 p. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** Psychiatric Management [A, A-, B, C, D, E, F, G] "Psychiatric management consists of a broad array of interventions and activities that psychiatrists should initiate and continue to provide to patients with major depressive disorder through all phases of treatment [I]." Acute Phase [A, A-, B, C, D, E, F, G] "Treatment in the acute phase should be aimed at inducing remission of the major depressive episode and achieving a full return to the patient's baseline level of functioning [I]. Acute phase treatment may include pharmacotherapy, depression-focused psychotherapy, the combination of medications and psychotherapy, or other somatic therapies such as

Systematic Review	Evidence
	electroconvulsive therapy (ECT), transcranial magnetic stimulation (TMS), or light therapy, as described in the sections that follow. Selection of an initial treatment modality should be influenced by clinical features (e.g., severity of symptoms, presence of co-occurring disorders or psychosocial stressors) as well as other factors (e.g., patient preference, prior treatment experiences) [I]. Any treatment should be integrated with psychiatric management and any other treatments being provided for other diagnoses [I]."
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline (see Section VI). Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial. [C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time.

Systematic Review	Evidence
	[E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis. [F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Provide all other grades and definitions from the evidence grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** The evidence base for practice guidelines is derived from two sources: research studies and clinical consensus. Where gaps exist in the research data, evidence is derived from clinical consensus, obtained through broad review of multiple drafts of each guideline (see Section VI). Both research data and clinical consensus vary in their validity and reliability for different clinical situations; guidelines state explicitly the nature of the supporting evidence for specific recommendations so that readers can make their own judgments regarding the utility of the recommendations. The following coding system is used for this purpose: [A] Randomized, double-blind clinical trial. A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments. [A–] Randomized clinical trial. Same as above but not double blind. [B] Clinical trial. A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial. [C] Cohort or longitudinal study. A study in which subjects are prospectively followed over time without any specific intervention. [D] Control study. A study in which a group of patients and a group of control subjects are identified in the present and information about them is pursued retrospectively or backward in time. [E] Review with secondary data analysis. A structured analytic review of existing data, e.g., a meta-analysis or a decision analysis.

Systematic Review	Evidence
	[F] Review. A qualitative review and discussion of previously published literature without a quantitative synthesis of the data. [G] Other. Opinion-like essays, case reports, and other reports not categorized above
Grade assigned to the **recommendation** with definition of the grade	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** [I] Recommended with substantial clinical confidence.
Provide all other grades and definitions from the recommendation grading system	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** Each recommendation is identified as falling into one of three categories of endorsement, indicated by a bracketed Roman numeral following the statement. The three categories represent varying levels of clinical confidence regarding the recommendation: [I] Recommended with substantial clinical confidence. [II] Recommended with moderate clinical confidence. [III] May be recommended on the basis of individual circumstances
Body of evidence: • Quantity – how many studies? • Quality – what type of studies?	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** Relevant updates to the literature were identified through a MEDLINE literature search for articles published since the second edition of the guideline, published in 2000. For this edition of the guideline, literature was identified through a computerized search of MEDLINE, using PubMed, for the period from January 1999 to December 2006. Using the MeSH headings depression or depressive disorder, as well as the key words major depression, major depressive disorder, neurotic depression, neurotic depressive, dysthymia, dysthymic, etc. yielded 39,157 citations. An additional 8,272 citations were identified by using the key words depression or depressive in combination with the MeSH headings affective disorders or psychotic or psychosis, psychotic, catatonic, catatonia, mood disorder, etc. This yielded 13,506 abstracts, which were screened for relevance with a very modest threshold for

Systematic Review	Evidence
	inclusion, then reviewed by the Work Group. The Psychoanalytic Electronic Publishing database (http://www.p-e-p.org) was also searched using the terms major depression or major depressive. This search yielded 112 references. The Cochrane databases were also searched for the key word depression, and 168 meta-analyses were identified. Additional, less formal, literature searches were conducted by APA staff and individual Work Group members and included references through May 2009. Sources of funding were considered when the Work Group reviewed the literature.
Estimates of benefit and consistency across studies	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** "The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline."
What harms were identified?	**2020 submission:** No updates. This guideline has not been updated. **2016 Submission:** "The literature review will include other guidelines addressing the same topic, when available. The work group constructs evidence tables to illustrate the data regarding risks and benefits for each treatment and to evaluate the quality of the data. These tables facilitate group discussion of the evidence and agreement on treatment recommendations before guideline text is written. Evidence tables do not appear in the guideline; however, they are retained by APA to document the development process in case queries are received and to inform revisions of the guideline."
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	**2020 submission:** No updates. This guideline has not been updated.

Systematic Review	Evidence
	2016 Submission: N/A Numerous (>100) studies related to follow-up for patients with mental illness have been published since the publication of this guideline, none of which contraindicate the need for appropriate follow-up after hospitalization for mental illness.

_____

1a.4 OTHER SOURCE OF EVIDENCE

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

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

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

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

1b. Performance Gap

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

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

0576_FUH_MEF_nqf_evidence_attachment_7.1.docx

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

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

Yes

1b. Performance Gap

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

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

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

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

This measure assesses whether health plan members who were hospitalized for a mental illness or intentional self-harm received a timely follow-up visit. Follow-up care following an acute event, such as hospitalization, reduces the risk of negative outcomes (e.g., medication errors, re-admission, emergency department use). Efforts to facilitate treatment following a hospital discharge also lead to less attrition in the initial post-acute period of treatment. Thus, this time period may be an important opportunity for health plans to implement strategies aimed at establishing strong relationships between patients and mental health providers and facilitate ongoing engagement in treatment.

According to an analysis of data from the National Inpatient Sample (NIS), between 2007 and 2014 there were over 1.5 million nonfatal suicide attempts requiring hospitalization, a rate of 67.1 per 100,000 persons (Connor et al., 2019). Another analysis of the NIS found that of 122,574 hospital discharges in 2003 with an injury diagnosis, 7.6% were for intentional self-harm (Patrick et al., 2010).

Fontanella et al. (2020) examined the association between timely outpatient follow-up after a psychiatric hospitalization and risk of death by suicide, and found that youths with a follow-up visit within 7 days of discharge had a significantly lower risk of death by suicide. A study of 90-day readmissions among individuals with schizophrenia and bipolar disorder found that individuals with an outpatient visit within 30-days following discharge experienced a lower risk of readmission within the following 90 days (Marcus et al., 2017). Similarly, Mark and colleagues (2013) found that increased follow-up at community mental health centers was associated with lower risk of re-admission among Medicaid patients hospitalized for mental illness or substance use disorder.

Evidence suggests that brief, low-intensity interventions are effective in bridging the gap between inpatient and outpatient treatment (Dixon 2009) and improving patient experience of continuity of care (Tomita & Herman, 2015). Low-intensity interventions are typically implemented at periods of high risk for treatment dropout, such as following an emergency room or hospital discharge or the time of entry into outpatient treatment. For example, Boyer et al evaluated strategies aimed at increasing attendance at outpatient appointments following hospital discharge. They found that the most common factor in a patient's medical history that was linked to a patient having a follow-up visit was a discussion about the discharge plan between the inpatient staff and outpatient clinicians. Other strategies they found that increased attendance at appointments included having the patient meet with outpatient staff and visit the outpatient program prior to discharge (Boyer 2000).

Barekatain M, Maracy MR, Rajabi F, Baratian H. (2014). Aftercare services for patients with severe mental disorder: A randomized controlled trial. J Res Med Sci. 19(3):240-5.

Boyer, C. A., McAlpine, D. D., Pottick, K. J., & Olfson, M. (2000). Identifying risk factors and key strategies in linkage to outpatient psychiatric care. The American journal of psychiatry, 157(10), 1592–1598. https://doi.org/10.1176/appi.ajp.157.10.1592

Conner, A., Azrael, D., & Miller, M. (2019). Suicide Case-Fatality Rates in the United States, 2007 to 2014: A Nationwide Population-Based Study. Annals of internal medicine, 171(12), 885–895. https://doi.org/10.7326/M19-1324

Dixon L, Goldberg R, Iannone V, et al. Use of a critical time intervention to promote continuity of care after psychiatric inpatient hospitalization for severe mental illness. Psychiatr Serv. 2009;60:451–458.

Fontanella, C. A., Warner, L. A., Steelesmith, D. L., Brock, G., Bridge, J. A., & Campo, J. V. (2020). Association of Timely Outpatient Mental Health Services for Youths After Psychiatric Hospitalization With Risk of Death by Suicide. JAMA network open, 3(8), e2012887.

Kreyenbuhl, J., Nossel, I., & Dixon, L. (2009). Disengagement from mental health treatment among individuals with schizophrenia and strategies for facilitating connections to care: A review of the literature. Schizophrenia Bulletin, 35, 696-703.

Luxton DD, June JD, Comtois KA. (2013). Can postdischarge follow-up contacts prevent suicide and suicidal behavior? A review of the evidence. Crisis. 34(1):32-41. doi: 10.1027/0227-5910/a000158.

Marcus, S. C., Chuang, C. C., Ng-Mak, D. S., & Olfson, M. (2017). Outpatient Follow-Up Care and Risk of Hospital Readmission in Schizophrenia and Bipolar Disorder. Psychiatric services (Washington, D.C.), 68(12), 1239–1246.

Mark, T., Tomic, K. S., Kowlessar, N., Chu, B. C., Vandivort-Warren, R., & Smith, S. (2013). Hospital Readmission Among Medicaid Patients with an Index Hospitalization for Mental and/or Substance Use Disorder. The Journal of Behavioral Health Services & Research, 40(2), 207–221.

Patrick, A. R., Miller, M., Barber, C. W., Wang, P. S., Canning, C. F., & Schneeweiss, S. (2010). Identification of hospitalizations for intentional self-harm when E-codes are incompletely recorded. Pharmacoepidemiology and drug safety, 19(12), 1263–1275. https://doi.org/10.1002/pds.2037

Tomita, A., & Herman, D. B. (2015). The role of a critical time intervention on the experience of continuity of care among persons with severe mental illness after hospital discharge. The Journal of nervous and mental disease, 203(1), 65–70.

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

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

HEDIS MY 2018 Variation in Performance across Health Plans- Commercial

Year	Rate	N	Mean	StDev	Min	P10	P25	Median	P75	P90	Max
2018	7-day rate	361	0.44	.11	0	0.3	0.37	0.44	0.51	0.59	0.79
	30-day rate	358	0.6	.11	0	0.52	0.6	0.67	0.73	0.78	0.92
2017	7-day rate	356	.46	.11	.18	.31	.38	.46	.54	.62	.78
	30-day rate	355	.68	.10	.38	.54	.62	.69	.75	.80	.91

HEDIS MY 2018 Variation in Performance across Health Plans- Medicare

Year	Rate	N	Mean	St Dev	Min	P10	P25	Median	P75	P90	Max
2018	7-day rate	308	0.28	.13	0	0.13	0.18	0.25	0.34	0.46	0.68
	30-day rate	308	0.48	.15	0.07	0.3	0.37	0.47	0.6	0.7	0.84
2017	7-day rate	304	.32	.13	.02	.18	.23	.29	.40	.50	.80
	30-day rate	304	.53	.15	.05	.35	.42	.52	.65	.74	.93

HEDIS MY 2018 Variation in Performance across Health Plans- Medicaid

Year	Rate	N	Mean	St Dev	Min	P10	P25	Median	P75	P90	Max

2018	7-day rate	173	0.36	.12	0.05	0.21	0.29	0.35	0.43	0.52	0.7
	30-day rate	172	0.57	.13	0.12	0.38	0.5	0.58	0.66	0.72	0.83
2017	7-day rate	183	.37	.13	.00	.19	.30	.37	.46	.54	.74
	30-day rate	183	.58	.14	.05	.40	.50	.60	.68	.74	.90

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

N/A

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

HEDIS data are stratified by type of insurance (e.g., Commercial, Medicaid, Medicare). While not specified in the measure, this measure can also be stratified by demographic variables, such as race/ethnicity or socioeconomic status, in order to assess the presence of health care disparities if the data are available to a plan. NCQA is actively engaged with partners including the CMS Office of Minority Health in identifying feasible methods to further integrate social risk factors into health plan quality measures, with a focus on stratification. Our work is aligned with recent recommendations from MedPAC and ASPE on optimal methods for addressing social risk in quality measurement and programs.1,2 This is an NCQA wide initiative. Our intent is to implement methods to bridge data concerns in the future.

HEDIS includes two measures that can be used as tools for assessing race/ethnicity and language needs of a plan's population: Race/Ethnicity Diversity of Membership and the Language Diversity of Membership. These measures promote standardized methods for collecting these data and follow Office of Management and Budget and National Academy of Medicine guidance 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.

1. Medicare Payment Advisory Commission. (2020). The Medicare Advantage program: Status report. In Report to the Congress: Medicare Payment Policy (p. 397). http://medpac.gov/docs/default-source/reports/mar20_medpac_ch13_sec.pdf
2. Office of the Assistant Secretary for Planning and Evaluation, & U.S. Department of Health & Human Services. (2020). Second Report to Congress on Social Risk and Medicare's Value-Based Purchasing Programs. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-basedpurchasing-programs

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

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Behavioral Health

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

Care Coordination, Safety

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

Children, Elderly, Populations at Risk

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

NA

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

This is not an eMeasure **Attachment:**

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

Attachment **:** 0576_FUH_Fall_2020_Value_Sets.xlsx

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

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

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

Not an instrument-based measure

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

Yes

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

Summary of most significant changes since previous submission:

- Since the last submission, several important changes were made to the measure:
- Added telehealth to the measure numerators
- The numerator was revised to no longer include visits that occur on the date of discharge. This change was made because an encounter on the date of discharge after hospitalization should be viewed as an intervention designed to support the patient and improve his or her likelihood of receiving timely follow-up care. Visits on the date of discharge should not be the only follow-up that patients receive and would not be considered good quality of care on their own; therefore, they do not meet the intent of the measure.

- The denominator was revised to include members with a principal diagnosis of intentional self-harm. This change was made to ensure that patients who are hospitalized for intentional self-harm are included in the measure because they warrant follow-up care, even if an accompanying mental health diagnosis is not present on the discharge claim.
- Expanded the definition of mental health provider to include Community Mental Health Centers (CMHC) and Certified Community Behavioral Health Clinics (CCBHC).

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

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

30-Day Follow-Up: A follow-up visit with a mental health provider within 30 days after discharge.

7-Day Follow-Up: A follow-up visit with a mental health provider within 7 days after discharge.

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

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

For both indicators, 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 mental health provider.
- An outpatient visit (BH Outpatient Value Set) with a mental health provider.
- An intensive outpatient encounter or partial hospitalization (Visit Setting Unspecified Value Set) with (Partial Hospitalization POS Value Set).
- An intensive outpatient encounter or partial hospitalization (Partial Hospitalization or Intensive Outpatient Value Set).
- A community mental health center visit (Visit Setting Unspecified Value Set; BH Outpatient Value Set; Observation Value Set; Transitional Care Management Services Value Set) with (Community Mental Health Center POS 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).
- A telehealth visit: (Visit Setting Unspecified Value Set) with (Telehealth POS Value Set) with a mental health provider.
- An observation visit (Observation Value Set) with a mental health provider.
- Transitional care management services (Transitional Care Management Services Value Set), with a mental health provider.
- A visit in a behavioral healthcare setting (Behavioral Healthcare Setting Value Set).
- A telephone visit (Telephone Visits Value Set) with a mental health provider.

(See corresponding Excel document for the value sets referenced above).

Mental Health Provider Definition:

A provider who delivers mental health services and meets any of the following criteria:

- An MD or doctor of osteopathy (DO) who is certified as a psychiatrist or child psychiatrist by the American Medical Specialties Board of Psychiatry and Neurology or by the American Osteopathic Board of Neurology and Psychiatry; or, if not certified, who successfully completed an accredited program of graduate medical or osteopathic education in psychiatry or child psychiatry and is licensed to practice patient care psychiatry or child psychiatry, if required by the state of practice.
- An individual who is licensed as a psychologist in his/her state of practice, if required by the state of practice.
- An individual who is certified in clinical social work by the American Board of Examiners; who is listed on the National Association of Social Worker's Clinical Register; or who has a master's degree in social work and is licensed or certified to practice as a social worker, if required by the state of practice.
- A registered nurse (RN) who is certified by the American Nurses Credentialing Center (a subsidiary of the American Nurses Association) as a psychiatric nurse or mental health clinical nurse specialist, or who has a master's degree in nursing with a specialization in psychiatric/mental health and two years of supervised clinical experience and is licensed to practice as a psychiatric or mental health nurse, if required by the state of practice.
- An individual (normally with a master's or a doctoral degree in marital and family therapy and at least two years of supervised clinical experience) who is practicing as a marital and family therapist and is licensed or a certified counselor by the state of practice, or if licensure or certification is not required by the state of practice, who is eligible for clinical membership in the American Association for Marriage and Family Therapy.
- An individual (normally with a master's or doctoral degree in counseling and at least two years of supervised clinical experience) who is practicing as a professional counselor and who is licensed or certified to do so by the state of practice, or if licensure or certification is not required by the state of practice, is a National Certified Counselor with a Specialty Certification in Clinical Mental Health Counseling from the National Board for Certified Counselors (NBCC).
- A physician assistant who is certified by the National Commission on Certification of Physician Assistants to practice psychiatry.
- A certified Community Mental Health Center (CMHC), or the comparable term (e.g. behavioral health organization, mental health agency, behavioral health agency) used within the state in which it is located, or a Certified Community Behavioral Health Clinic (CCBHC).
- Only authorized CMHCs are considered mental health providers. To be authorized as a CMHC, an entity must meet one of the following criteria:
- The entity has been certified by CMS to meet the conditions of participation (CoPs) that community mental health centers (CMHCs) must meet in order to participate in the Medicare program, as defined in the Code of Federal Regulations Title 42. CMS defines a CMHC as an entity that meets applicable licensing or certification requirements for CMHCs in the State in which it is located and provides the set of services specified in section 1913(c)(1) of the Public Health Service Act (PHS Act).
- The entity has been licensed, operated, authorized, or otherwise recognized as a CMHC by a state or county in which it is located.
- Only authorized CCBHCs are considered mental health providers. To be authorized as a CCBHC, an entity must meet one of the following criteria:

  - Has been certified by a State Medicaid agency as meeting criteria established by the Secretary for participation in the Medicaid CCBHC demonstration program pursuant to Protecting Access to Medicare Act § 223(a) (42 U.S.C. § 1396a note); or as meeting criteria within the State's Medicaid Plan to be considered a CCBHC.

- Has been recognized by the Substance Abuse and Mental Health Services Administration, through the award of grant funds or otherwise, as a CCBHC that meets the certification criteria of a CCBHC.

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

Discharges from an acute inpatient setting with a principal diagnosis of mental illness or intentional self-harm on the discharge claim during the first 11 months of the measurement year (i.e. January 1 to December 1) for members 6 years and older.

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

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

An acute inpatient discharge with a principal diagnosis of mental illness or intentional self-harm (Mental Illness Value Set; Intentional Self-Harm Value Set) on the discharge claim on or between January 1 and December 1 of the measurement year. To identify acute inpatient discharges:

1. Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
2. Exclude nonacute inpatient stays (Nonacute Inpatient Stay Value Set).
3. Identify the discharge date for the stay.

The denominator for this measure is based on discharges, not on members. If members have more than one discharge, include all discharges on or between January 1 and December 1 of the measurement year.

Acute readmission or direct transfer

Identify readmissions and direct transfers to an acute inpatient care setting during the 30-day follow-up period:

- Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
- Exclude nonacute inpatient stays (Nonacute Inpatient Stay Value Set).
- Identify the admission date for the stay.

Exclude both the initial discharge and the readmission/direct transfer discharge if the last discharge occurs after December 1 of the measurement year.

If the readmission/direct transfer to the acute inpatient care setting was for a principal diagnosis (use only the principal diagnosis on the discharge claim) of mental health disorder or intentional self-harm (Mental Health Diagnosis Value Set; Intentional Self-Harm Value Set), count only the last discharge.

If the readmission/direct transfer to the acute inpatient care setting was for any other principal diagnosis (use only the principal diagnosis on the discharge claim) exclude both the original and the readmission/direct transfer discharge.

See corresponding Excel document for the Value Sets referenced above in S.2b.

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

Exclude from the denominator for both rates, members who begin using hospice services anytime during the measurement year (Hospice Value Set)

Exclude both the initial discharge and the readmission/direct transfer discharge if the readmission/direct transfer discharge occurs after December 1 of the measurement year.

Exclude discharges followed by readmission or direct transfer to a nonacute facility within the 30-day follow-up period regardless of principal diagnosis.

Exclude discharges followed by readmission or direct transfer to an acute facility within the 30-day follow-up period if the principal diagnosis was not for mental health or intentional self-harm.

These discharges are excluded from the measure because rehospitalization or transfer may prevent an outpatient follow-up visit from taking place.

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

Members in hospice are excluded from the eligible population.

Exclude both the initial discharge and the readmission/direct transfer discharge if the last discharge occurs after December 1 of the measurement year.

If the readmission/direct transfer to the acute inpatient care setting was for a principal diagnosis (use only the principal diagnosis on the discharge claim) of mental health disorder or intentional self-harm (Mental Health Diagnosis Value Set; Intentional Self-Harm Value Set), count only the last discharge.

If the readmission/direct transfer to the acute inpatient care setting was for any other principal diagnosis (use only the principal diagnosis on the discharge claim) exclude both the original and the readmission/direct transfer discharge

Exclude discharges followed by readmission or direct transfer to a nonacute inpatient care setting within the 30-day follow-up period, regardless of principal diagnosis for the readmission. To identify readmissions and direct transfers to a nonacute inpatient care setting:

- Identify all acute and nonacute inpatient stays (Inpatient Stay Value Set).
- Confirm the stay was for nonacute care based on the presence of a nonacute code (Nonacute Inpatient Stay Value Set) on the claim.
- Identify the admission date for the stay.

These discharges are excluded from the measure because rehospitalization or direct transfer may prevent an outpatient follow-up visit from taking place.

See corresponding Excel document for the Value Sets referenced above in S.2b.

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

N/A

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

No risk adjustment or risk stratification

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Higher score

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

Step 1. Determine the denominator. The denominator is all discharges that meet the specified denominator criteria (S7).

Step 2. Remove exclusions. Remove all discharges from the denominator that meet the specified exclusion criteria (S9).

Step 3. Identify numerator events: Search administrative systems to identify numerator events for all discharges in the denominator (S5).

Step 4. Calculate the rate by dividing the events in step 3 by the discharges in step 2.

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

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

N/A

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

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

N/A

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

If other, please describe in S.18.

Claims

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

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

This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from Health 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)*

Inpatient/Hospital, Outpatient Services

If other:

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

N/A

2. Validity – See attached Measure Testing Submission Form

nqf_testing_attachment_7.1_2020-637395029781340613.docx

2.1 For maintenance of endorsement

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

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 For maintenance of endorsement

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

No - This measure is not risk-adjusted

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

Measure Number (*if previously endorsed*)**:** 0576
Measure Title: Follow-Up After Hospitalization for Mental Illness
Date of Submission: 8/3/2020

Type of Measure:

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

*cell intentionally left blank

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

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

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

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

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

2020 Submission:

N/A

2016 Submission:

N/A

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

2020 Submission

Testing of measure score reliability and validity was performed using data from calendar year 2018.

2016 Submission:

2009-2011

2014-2016

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

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

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the*

analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)

2020 Submission:

This measure assesses the percentage of discharges for members 6 years of age and older who were hospitalized for treatment of selected mental illness or intentional self-harm diagnoses and who had a follow-up visit with a mental health provider. Two rates are reported:

1. The percentage of discharges for which the member received follow-up within 30 days after discharge.
2. The percentage of discharges for which the member received follow-up within 7 days after discharge.

Testing was completed at the health plan level which is appropriate for the level of reporting for this measure.

Measure score reliability testing and construct validity testing: The measure score reliability was calculated from HEDIS data that included 358 Commercial health plans, 172 Medicaid plans, and 308 Medicare plans. The sample included all Commercial, Medicare and Medicaid health plans submitting data to NCQA for this HEDIS measure. The plans were geographically diverse and varied in size.

2016 Update: MEASURE SCORE RELIABILITY TESTING

MEASURE SCORE RELIABILITY TESTING

The measure score reliability was calculated from 2016 HEDIS data that included 368 Commercial health plans, 166 Medicaid health plans, and 301 Medicare health plans for the 7-day follow-up rate and 368 Commercial health plans, 168 Medicaid health plans, and 301 Medicare health plans for the 30-day follow-up rate. The sample included all health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

SYSTEMATIC EVALUATION OF FACE VALIDITY

The Follow-Up After Hospitalization for Mental Illness measure was tested for face validity with several panels of experts. Measurement Advisory Panels (MAP) provide the clinical and technical knowledge required to develop the measures. The Behavioral Health MAP included 12 experts in behavioral health including representation by consumers, health plans, health care providers and policy makers. NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 15 members. The CPM is organized and managed by NCQA, and is responsible for advising NCQA staff on the development and maintenance of performance measures. The CPM also meets with the NCQA Board of Directors to recommend measures for inclusion in HEDIS. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement. Additional HEDIS Expert Panels provide invaluable assistance by identifying methodological issues and giving feedback on new and existing measures. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel.

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

2020 Submission

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 average eligible population for the measure across health plans. For this measure, the eligible population is the number of eligible discharges among plan members 6 years of age and older.

7-day Follow-Up Rate

Product Line	Number of Plans	Mean number of eligible discharges per plan
Commercial	361	668
Medicaid	173	1946
Medicare	308	344

30-Day Follow-Up Rate

Product Line	Number of Plans	Mean number of eligible discharges per plan
Commercial	358	665
Medicaid	172	1956
Medicare	308	344

2016 Update: MEASURE SCORE RELIABILITY TESTING

Patients included for measure score reliability testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare beneficiaries. Data are summarized at the health plan level and stratified by product line. Below is a description of the testing data, including number of health plans included and the mean eligible population for the measure across health plans.

7-day Follow-Up Rate

Product Line	Number of Plans	Mean number of eligible patients per plan
Commercial	368	568
Medicaid	166	1,182
Medicare	301	279

30-Day Follow-Up Rate

Product Line	Number of Plans	Mean number of eligible patients per plan
Commercial	368	568
Medicaid	168	1,169
Medicare	301	279

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

2020 Submission
No differences in the data used for reliability and construct validity testing.

2016 Update: MEASURE SCORE RELIABILITY TESTING

Reliability of the measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data for the measure (described above).

Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions we have described the composition of the technical expert panel which assessed face validity in the data sample questions above.

1.8 What were the 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.

2020 Submission
We did not assess data by social risk factors. Social risk factor data were not available in reported results. This measure is specified for Medicare, Medicaid and Commercial members aged 6 and older. NCQA is actively engaged with partners including the CMS Office of Minority Health in identifying feasible methods to further integrate social risk factors into health plan quality measures, with a focus on stratification. This is aligned with recent recommendations from MedPAC and ASPE on optimal methods for addressing social risk in quality measurement and programs.1,2This is an NCQA wide initiative. Our intent is to implement methods to bridge data concerns in the future.

1. Medicare Payment Advisory Commission. (2020). The Medicare Advantage program: Status report. In Report to the Congress: Medicare Payment Policy (p. 397). http://medpac.gov/docs/default-source/reports/mar20_medpac_ch13_sec.pdf
2. Office of the Assistant Secretary for Planning and Evaluation, & U.S. Department of Health & Human Services. (2020). Second Report to Congress on Social Risk and Medicare's Value-Based Purchasing Programs. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-basedpurchasing-programs

2016 Update

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

2012 Submission

The measure is not stratified to detect disparities. NCQA has participated with IOM and others in attempting to include information on disparities in measure data collection. However, at the present time, this data, at all

levels (claims data, paper chart review, and electronic records), is not coded in a standard manner, and is incompletely captured. There are no consistent standards for what entity (physician, group, plan, employer) should capture and report this data. While "requiring" reporting of the data could push the field forward, it has been our position that doing so would create substantial burden with inability to use the data because of its inconsistency. At the present time, we agree with the IOM report that disparities are best considered by the use of zip code analysis which has limited applicability in most reporting situations. At the health plan level, for HEDIS health plan data collection, NCQA does have extensive data related to our use of stratification by insurance status (Medicare, Medicaid and private-commercial) and would strongly recommend this process where the data base supporting the measurement includes this information. However, we believe that the measure specifications should NOT require this since the measure is still useful where the data needed to determine disparities cannot be ascertained from the data available.

2a2. RELIABILITY TESTING

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

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

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

2020 Submission

Reliability testing of performance measure score

We utilized the methodology described by John Adams (Adams, J.L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009) to calculate signal-to-noise reliability. This methodology uses the Beta-binomial model to assess how well one can confidently distinguish the performance of one reporting 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 across reporting entities (plans, physicians, etc.) in performance. The Beta-binomial model is an appropriate model when estimating the reliability of simple pass/fail rate measures, such as the Follow-Up After Hospitalization for Mental Illness measure. 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 reporting entities.

For the Follow-Up After Hospitalization for Mental Illness measure, health plans are the reporting entity. For the formulas and explanations below, we use health plans as the reporting entity.

The formula for signal-to-noise reliability is:

$$\text{Signal-to-noise reliability} = \sigma^2_{\text{plan-to-plan}} / (\sigma^2_{\text{plan-to-plan}} + \sigma^2_{\text{error}})$$

Therefore, we need to estimate two variances: 1) variance between plans ($\sigma^2_{\text{plan-to-plan}}$); 2) variance within plans (σ^2_{error}).

1. Variance between plans = $\sigma^2_{\text{plan-to-plan}} = (\alpha\ \beta) / (\alpha + \beta + 1)(\alpha + \beta)^2$

α and β are two shape parameters of the Beta-Binomial distribution, α >0, β > 0

2. Variance within plans: $\sigma^2_{error} = \hat{p}(1- \hat{p})/n$

   $\hat{p}$ = observed rate for the plan

   n = plan-specific denominator for the observed rate (most often the number of eligible plan members; in this case, the number of eligible discharges associated with each plan)

Using Adams' 2009 methodology, we estimated the reliability for each reporting entity, then averaged these reliability estimates across all reporting entities to produce a point estimate of signal-to-noise reliability. We label this point estimate "mean signal-to-noise reliability". The mean signal-to-noise reliability measures how well, on average, the measure can differentiate between reporting entity performance on the measure.

Along with the point estimate of mean signal-to-noise reliability, we are also providing:

1. The standard error (SE) and 95% confidence interval (95% CI) of the mean signal-to-noise reliability for all plans and stratified by the denominator size (number of eligible members per plan). The SE and 95% CI of the mean signal-to-noise reliability provides information about the stability of reliability. The 95% CI is the mean signal-to-noise reliability ± (1.96*SE). We also stratified the results by the denominator size using terciles of the distribution to provide additional information about the stability of reliability.

2. The distribution (minimum, 10th, 25th, 50th, 75th, 90th, maximum) of the plan-level signal-to-noise reliability estimates. Each plan's reliability estimate is a ratio of signal to noise, as described above [$\sigma^2_{plan\text{-}to\text{-}plan} / (\sigma^2_{plan\text{-}to\text{-}plan} + \sigma^2_{error})$]. Variability between plans ($\sigma^2_{plan\text{-}to\text{-}plan}$) is the same for each plan, while the specific plan error (σ^2_{error}) varies. Reliability for each plan is an ordinal measure of how well one can determine where a plan lies in the distribution of reliability across all plans, with higher estimates indicating better reliability. We also stratified the results by the denominator size using terciles of the distribution to provide additional information about the distribution of plan-level signal-to-noise reliability estimates. The number of plans in each stratum and the per-plan denominators of indicators are displayed in the summary tables.

This methodology allows us to estimate the reliability for each plan and summarize the distribution of these estimates.

2016 Update: METHOD FOR MEASURE SCORE RELIABILITY TESTING METHOD FOR BETA-BINOMIAL RELIABILITY TESTING

The beta-binomial method (Adams, 2009) measures the proportion of total variation attributable to a health plan, which represents the *signal*. The beta-binomial model also estimates the proportion of variation attributable to measurement error for each plan, which represents *noise*. The reliability of the measure is represented as the ratio of signal to noise.

- A score of 0 indicates none of the variation (signal) is attributable to the plan
- A score of 1.0 indicates all of the variation (signal) is attributable to the plan
- A score of 0.7 or higher indicates adequate reliability to distinguish performance between two plans

PLAN-LEVEL RELIABILITY

The underlying formulas for the beta-binomial reliability can be adapted to construct a plan-specific estimate of reliability by substituting variation in the individual plan's variation for the average plan's variation. The reliability for some plans may be more or less than the overall reliability across plans.

Adams JL. The Reliability of Provider Profiling: A Tutorial. Santa Monica, CA: RAND Corp. TR-653-NCQA, 2009

2012 Submission

In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009) in work produced for the National Committee for Quality Assurance (NCQA).

The following is quoted from the tutorial which focused on provider-level assessment: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient." This approach is also relevant to health plans and other accountable entities.

The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Adams' approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures.

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

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

2020 Submission:

Signal-to-Noise Reliability Assessment for the *Follow-Up After Hospitalization for Mental Illness* Measure

Table 1 shows the point estimates of mean signal-to-noise reliability using above methodology.

Table 1. Point Estimates of Mean Signal-to-Noise Reliability by Product Type, Calendar Year 2018 Data

Follow-Up After Hospitalization for Mental Illness	**Point estimate: Mean Signal-To-Noise Reliability (Commercial)**	**Point estimate: Mean Signal-To-Noise Reliability (Medicaid)**	**Point estimate: Mean Signal-To-Noise Reliability (Medicare)**
Follow-Up After Hospitalization For Mental Illness - 7 days	0.884	0.969	0.900
Follow-Up After Hospitalization For Mental Illness - 30 days	0.883	0.967	0.910

Table 2a. Mean Signal-To-Noise Reliability, Standard Error (SE) and 95% Confidence Interval (95% CI) for the Follow-Up After Hospitalization For Mental Illness - (7 day) Measure by Terciles of the Denominator Size and for All Submissions Stratified by Plan Type, Calendar Year 2018 Data

Stratification	**Number of Plans**	**Number of Eligible Discharges per Plan (min - max)**	**Mean Signal-To-Noise Reliability**	**SE**	**95% CI**
All Commercial	**361**	**30-7412**	**0.884**	**0.006**	**(0.872, 0.895)**
Tercile 1	120	30-131	0.783	0.007	(0.769, 0.797)
Tercile 2	118	132-471	0.915	0.003	(0.910, 0.920)
Tercile 3	123	482-7412	0.976	0.001	(0.973, 0.978)
All Medicaid	**173**	**38-17406**	**0.969**	**0.003**	**(0.962, 0.975)**
Tercile 1	57	38-482	0.933	0.006	(0.921, 0.946)
Tercile 2	57	512-2030	0.987	0.001	(0.986, 0.989)
Tercile 3	59	2054-17406	0.994	0.000	(0.993, 0.994)
All Medicare	**308**	**30-4224**	**0.900**	**0.005**	**(0.890, 0.909)**
Tercile 1	102	30-90	0.815	0.007	(0.802, 0.828)
Tercile 2	101	91-270	0.920	0.003	(0.913, 0.927)
Tercile 3	105	273-4224	0.973	0.002	(0.970, 0.976)

SE: Standard Error of the mean.

95% CI: 95% confidence interval.

Table 2b. Mean Signal-To-Noise Reliability, Standard Error (SE) and 95% Confidence Interval (95% CI) for the Follow-Up After Hospitalization For Mental Illness - (30-day) Measure by Terciles of the Denominator Size and for All Submissions Stratified by Plan Type, Calendar Year 2018 Data

Stratification	Number of Plans	Number of Eligible Discharges per Plan (min - max)	Mean Signal-To-Noise Reliability	SE	95% CI
All Commercial	**361**	**30-7412**	**0.883**	**0.006**	**(0.872, 0.895)**
Tercile 1	120	30-131	0.807	0.007	(0.794, 0.821)
Tercile 2	116	132-470	0.921	0.002	(0.916, 0.926)
Tercile 3	122	471-7412	0.971	0.001	(0.969, 0.974)
All Medicaid	**174**	**38-17406**	**0.967**	**0.004**	**(0.960, 0.975)**
Tercile 1	57	38-512	0.932	0.007	(0.919, 0.945)
Tercile 2	56	529-2030	0.988	0.001	(0.986, 0.989)
Tercile 3	59	2054-17406	0.994	0.000	(0.993, 0.994)
All Medicare	**308**	**30-4224**	**0.910**	**0.004**	**(0.902, 0.918)**
Tercile 1	102	30-90	0.838	0.004	(0.829, 0.847)
Tercile 2	101	91-270	0.933	0.002	(0.929, 0.937)
Tercile 3	105	273-4224	0.974	0.001	(0.971, 0.977)

SE: Standard Error of the mean.

95% CI: 95% confidence interval.

Table 3a. Distribution of Plan-Level Signal-To-Noise Reliability for the Follow-Up after Hospitalization for Mental Illness 7-day measure rate by Terciles of the Denominator Size and for All Submissions by Plan Type, Calendar Year 2018 Data

Stratification	Number of Plans	Distribution of Plan Estimates of Signal-to-Noise Reliability: Min	Distribution of Plan Estimates of Signal-to-Noise Reliability: P10	Distribution of Plan Estimates of Signal-to-Noise Reliability: P25	Distribution of Plan Estimates of Signal-to-Noise Reliability: P50	Distribution of Plan Estimates of Signal-to-Noise Reliability: P75	Distribution of Plan Estimates of Signal-to-Noise Reliability: P90	Distribution of Plan Estimates of Signal-to-Noise Reliability: Max
All Commercial	**361**	**0.567**	**0.708**	**0.826**	**0.884**	**0.972**	**0.987**	**1.000**
Tercile 1	120	0.608	0.670	0.735	0.783	0.849	0.865	1.000
Tercile 2	118	0.852	0.876	0.894	0.915	0.941	0.948	0.959

Stratification	Number of Plans	Distribution of Plan Estimates of Signal-to-Noise Reliability: Min	Distribution of Plan Estimates of Signal-to-Noise Reliability: P10	Distribution of Plan Estimates of Signal-to-Noise Reliability: P25	Distribution of Plan Estimates of Signal-to-Noise Reliability: P50	Distribution of Plan Estimates of Signal-to-Noise Reliability: P75	Distribution of Plan Estimates of Signal-to-Noise Reliability: P90	Distribution of Plan Estimates of Signal-to-Noise Reliability: Max
Tercile 3	123	0.947	0.957	0.966	0.976	0.986	0.992	0.997
All Medicaid	**173**	**0.743**	**0.916**	**0.964**	**0.969**	**0.995**	**0.997**	**0.999**
Tercile 1	57	0.770	0.867	0.906	0.933	0.969	0.973	0.985
Tercile 2	57	0.976	0.979	0.983	0.987	0.991	0.993	0.997
Tercile 3	59	0.988	0.991	0.992	0.994	0.996	0.997	0.999
All Medicare	**308**	**0.653**	**0.771**	**0.848**	**0.900**	**0.967**	**0.987**	**1.000**
Tercile 1	102	0.667	0.721	0.776	0.815	0.859	0.902	1.000
Tercile 2	101	0.845	0.869	0.893	0.920	0.944	0.964	0.985
Tercile 3	105	0.939	0.952	0.958	0.973	0.988	0.993	0.997

Table 3b. Distribution of Plan-Level Signal-To-Noise Reliability for the Follow-Up after Hospitalization for Mental Illness 30-day measure rate by Terciles of the Denominator Size and for All Submissions by Plan Type, Calendar Year 2018 Data

Stratification	Number of Plans	Distribution of Plan Estimates of Signal-to-Noise Reliability: Min	Distribution of Plan Estimates of Signal-to-Noise Reliability: P10	Distribution of Plan Estimates of Signal-to-Noise Reliability: P25	Distribution of Plan Estimates of Signal-to-Noise Reliability: P50	Distribution of Plan Estimates of Signal-to-Noise Reliability: P75	Distribution of Plan Estimates of Signal-to-Noise Reliability: P90	Distribution of Plan Estimates of Signal-to-Noise Reliability: Max
All Commercial	**358**	**0.541**	**0.701**	**0.821**	**0.883**	**0.973**	**0.988**	**1.000**
Tercile 1	120	0.622	0.702	0.757	0.807	0.865	0.891	1.000
Tercile 2	116	0.846	0.881	0.903	0.921	0.943	0.951	0.968
Tercile 3	122	0.935	0.949	0.960	0.971	0.984	0.990	0.997
All Medicaid	**172**	**0.708**	**0.908**	**0.963**	**0.967**	**0.995**	**0.997**	**0.999**
Tercile 1	57	0.747	0.860	0.910	0.932	0.969	0.975	0.983
Tercile 2	56	0.975	0.979	0.984	0.988	0.992	0.994	0.995
Tercile 3	59	0.989	0.991	0.992	0.994	0.996	0.997	0.999
All Medicare	**308**	**0.714**	**0.788**	**0.861**	**0.910**	**0.967**	**0.989**	**0.997**

Stratification	Number of Plans	Distribution of Plan Estimates of Signal-to-Noise Reliability: Min	Distribution of Plan Estimates of Signal-to-Noise Reliability: P10	Distribution of Plan Estimates of Signal-to-Noise Reliability: P25	Distribution of Plan Estimates of Signal-to-Noise Reliability: P50	Distribution of Plan Estimates of Signal-to-Noise Reliability: P75	Distribution of Plan Estimates of Signal-to-Noise Reliability: P90	Distribution of Plan Estimates of Signal-to-Noise Reliability: Max
Tercile 1	102	0.737	0.779	0.801	0.838	0.874	0.893	0.944
Tercile 2	101	0.891	0.905	0.916	0.933	0.949	0.958	0.968
Tercile 3	105	0.949	0.955	0.961	0.974	0.988	0.992	0.997

2016 Update: MEASURE SCORE RELIABILITY

MEASURE LEVEL RELIABILITY

NCQA pools data reported by health plans according to product line. The mean reliability for the 7-day Rate per the beta binomial model was 0.97 for Commercial health plans, 0.96 for Medicare, and 0.99 for Medicaid. The mean reliability for the 30-day Rate was 0.96 for Commercial health plans, 0.97 for Medicare, and 0.99 for Medicaid.

Beta-Binomial Statistic For Each Measure Rate

Rate	*Commercial:* Avg	*Commercial:* Minimum	*Medicaid:* Avg	*Medicaid:* Minimum	*Medicare:* Avg	*Medicare:* Minimum
7-Day Follow-Up	1.0	0.7	1.0	0.8	1.0	0.7
30-Day Follow-Up	1.0	0.6	1.0	0.8	1.0	0.8

2012 Submission

Rate 1. The percentage of members who received follow-up within 30 days of discharge

Commercial: 0.967434

Medicaid: 0.988749

Medicare: 0.949915

Rate 2. The percentage of members who received follow-up within 7 days of discharge.

Commercial: 0.954861

Medicaid: 0.989110

Medicare: 0.951935

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

2020 submission:

In general, a score of 0.7 or higher suggests the measure has adequate reliability. The results suggest the measure has high reliability and more details are discussed below.

Table 2a provides the point estimate of mean signal-to-noise reliability, its standard error, and the 95% CI for the Follow-Up After Hospitalization For Mental Illness - 7 day measure rate for Commercial, Medicaid and Medicare plans overall and stratified by the denominator size (distribution of the number of eligible members per plan). Over all commercial plans, the reliability estimate is 0.884, and the 95% CI is (0.872, 0.895), indicating good reliability. Stratified analyses show that reliability increase as plan size gets larger and exceeds .8 for all terciles. Over all Medicaid plans, the reliability estimate is 0.969 and the 95% CI is (0.962, 0.975), indicating very good reliability. Results from the stratified analyses show that reliability exceeds 0.9 for all terciles. Over all Medicare plans, the reliability estimate is 0.900 and the 95% CI is (0.890, 0.909), indicating very good reliability. Results from the stratified analyses show that reliability tends to increase as plan size gets larger and exceeds .8 for all terciles.

Table 2b provides the point estimate of mean signal-to-noise reliability, its standard error, and the 95% CI for the Follow-Up After Hospitalization For Mental Illness - 30 day measure rate for Commercial, Medicaid and Medicare plans overall and stratified by the denominator size (distribution of the number of eligible members per plan). Across all commercial plans, the reliability estimate is 0.883, and the 95% CI is (0.872, 0.895) indicating good reliability. Stratified analyses show that reliability increases as plan size gets larger and exceeds .8 for all terciles. Across all Medicaid plans, the reliability estimate is 0.967 and the 95% CI is (0.960, 0.975), indicating very good reliability. Results from the stratified analyses show that reliability exceeds 0.9 for all terciles. Across all Medicare plans, the reliability estimate is 0.910 and the 95% CI is (0.902, 0.918) indicating very good reliability. Results from the stratified analyses show that reliability increases as plan size gets larger and exceeds 0.8 for all terciles.

Table 3a summarizes the distribution of plan-level signal-to-noise reliability estimates for the Follow-Up After Hospitalization for Mental Illness 7-day measure rate. Over all commercial plans, the estimates range from 0.567 to 0.100. The 50th percentile is 0.884, which exceeds the 0.70 threshold for reliability. For Medicaid plans, the estimates range from 0.743 to .999; the 10th percentile is 0.916, indicating very good reliability. For Medicare plans, the estimates range from 0.653 to 1.000; the 50th percentile is 0.900, which exceeds the 0.70 threshold for reliability. This table also include the distribution of plan-level signal-to-noise reliability estimates stratified by denominator size. Reliability estimates tend to be higher for plans with a larger denominator.

Table 3b summarizes the distribution of plan-level signal-to-noise reliability estimates for the Follow-Up After Hospitalization for Mental Illness 30-day measure rate. Over all commercial plans, the estimates range from 0.541 to 1.000. The 50th percentile is 0.883, which exceeds the 0.70 threshold for reliability. For Medicaid plans, the estimates range from 0.708 to 0.999; the 10th percentile is 0.908, indicating very good reliability. For Medicare plans, the estimates range from 0.714 to 0.997; the 50th percentile is 0.910, which exceeds the 0.70 threshold for reliability. This table also include the distribution of plan-level signal-to-noise reliability estimates stratified by denominator size. Reliability estimates tend to me higher for plans with a larger denominator.

2016 submission:

Among Commercial, Medicare, and Medicaid plans, results indicate both the 7-day and 30-day rates within this measure have a good signal to noise ratio that are well above the 0.7 threshold for adequate reliability. This data analysis suggests the measure has high reliability and can discriminate performance between accountable entities.

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)*

2020 Submission

We assessed construct validity and face 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 Hospitalization for Mental Illness* measure positively correlated with one another?
- Is the *Follow-Up After Hospitalization For Mental Illness* measure positively correlated with the HEDIS *Follow-Up After Emergency Department Visit for Mental Illness* measure which assesses emergency department (ED) visits for adults and children 6 years of age and older with a diagnosis of mental illness and who received a follow-up visit for mental illness within 7- and 30-days?

We hypothesized that rates within the *Follow-Up After Hospitalization For Mental Illness* measure would be highly positively correlated, and that organizations that perform well on *Follow-Up After Hospitalization For Mental Illness* should perform well on the other measure, *Follow-Up After Emergency Department Visit for Mental Illness*, given that they address the same or similar populations and that they address similar activities for patients following an acute event involving mental illness.

NCQA performs Pearson correlation for construct validity using HEDIS health plan data. The test estimates the strength of linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a strong positive linear association: an increase in values of one variable is associated with increase in value of another variable. A value of 0 indicates no linear association. A value of -1 indicates a strong negative relationship in which an increase in values of the first variable is associated with a decrease in values of the second variable. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The sample size for the correlation analysis is the number of plans that reported both measures. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We adjusted our p-values to account for testing multiple correlations and used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of assessing face validity

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, including the Behavioral Health Measurement Advisory Panel (BHMAP), Geriatric Measurement Advisory Panel (GMAP), 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.

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.

2016 Update

Method of Assessing Face Validity

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

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

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

uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.

NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

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

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

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

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

ICD-10 CONVERSION

The below steps describe our methods to convert this measure to ICD-10 in order to develop a new code set fully consistent with the intent of the measure.

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use General Equivalence Mapping (GEM) to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.

4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

Tools Used to Identify/Map to ICD-10
All tools used for mapping/code identification from CMS ICD-10 website (https://www.cms.gov/medicare/Coding/ICD10/index.html).
GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

Expert Participation

The NCQA HEDIS Expert Coding Panel and NCQA's Behavioral Health Measurement Advisory Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

2012 Submission

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

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

NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure "Desirable"? The work-up is vetted by NCQA's MAPs, the TAG, the HEDIS Policy Panel and various other panels.

*Step 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase.

Development includes the following tasks.

1. Ensure funding throughout measure testing
2. Prepare a detailed conceptual and operational work-up that includes a testing proposal
3. Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures

The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

*Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to the CPM about new measures or about changes to existing measures.

NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

*Step 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's Quality Compass? Or in accreditation scoring.

The first-year distinction guarantees that a measure can be efficiently collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues.

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

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

Step 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reevaluated at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments contribute to measure evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

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

What makes a measure "Desirable"?

Whether considering the value of a new measure or the continuing worth of an existing one, we must define what makes a measure useful. HEDIS measures encourage improvement. The defining question for all performance measurement—"Where can measurement make a difference?"—can be answered only after considering many factors. NCQA has established three areas of desirable characteristics for HEDIS measures, discussed below.

30. Relevance: Measures should address features that apply to purchasers or consumers, or which will stimulate internal efforts toward quality improvement. More specifically, relevance includes the following attributes.

Meaningful: What is the significance of the measure to the different groups concerned with health care? Is the measure easily interpreted? Are the results meaningful to target audiences?

Measures should be meaningful to at least one HEDIS audience (e.g., individual consumers, purchasers or health care systems). Decision makers should be able to understand a measure's clinical and economic significance.

Important to health: What is the prevalence and overall impact of the condition in the U.S. population? What significant health care aspects will the measure address?

We should consider the type of measure (e.g., outcome or process), the prevalence of medical condition addressed by the measure and the seriousness of affected health outcomes.

Financially important: What financial implications result from actions evaluated by the measure? Does the measure relate to activities with high financial impact?

Measures should relate to activities that have high financial impact.

Cost effective: What is the cost benefit of implementing the change in the health care system? Does the measure encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness? Measures should encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness.

Strategically important: What are the policy implications? Does the measure encourage activities that use resources efficiently? Measures should encourage activities that use resources most efficiently to maximize member health.

Controllable: What impact can the organization have on the condition or disease? What impact can the organization have on the measure? Health care systems should be able to improve their performance. For outcome measures, at least one process should be controlled and have an important effect on outcome. For process measures, there should be a strong link between the process and desired outcome.

Variation across systems: Will there be variation across systems? There should be the potential for wide variation across systems.

Potential for improvement: Will organizations be able to improve performance? There should be substantial room for performance improvement.

31. Scientific soundness: Perhaps in no other industry is scientific soundness as important as in health care. Scientific soundness must be a core value of our health care system—a system that has extended and improved the lives of countless individuals.

Clinical evidence: Is there strong evidence to support the measure? Are there published guidelines for the condition? Do the guidelines discuss aspects of the measure? Does evidence document a link between clinical processes and outcomes addressed by the measure? There should be evidence documenting a link between clinical processes and outcomes.

Reproducible: Are results consistent? Measures should produce the same results when repeated in the same population and setting.

Valid: Does the measure make sense? Measures should make sense logically and clinically, and should correlate well with other measures of the same aspects of care.

Accurate: How well does the measure evaluate what is happening? Measures should precisely evaluate what is actually happening.

Risk adjustment: Is it appropriate to stratify the measure by age or another variable? Measure variables should not differ appreciably beyond the health care system's control, or variables should be known and measurable. Risk stratification or a validated model for calculating an adjusted result can be used for measures with confounding variables.

Comparability of data sources: How do different systems affect accuracy, reproducibility and validity? Accuracy, reproducibility and validity should not be affected if different systems use different data sources for a measure.

32. Feasibility:

The goal is not only to include feasible measures, but also to catalyze a process whereby relevant measures can be made feasible.

Precise specifications: Are there clear specifications for data sources and methods for data collection and reporting? Measures should have clear specifications for data sources and methods for data collection and reporting.

Reasonable cost: Does the measure impose a burden on health care systems? Measures should not impose an inappropriate burden on health care systems.

Confidentiality: Does data collection meet accepted standards of member confidentiality?

Data collection should not violate accepted standards of member confidentiality. Logistical feasibility

Are the required data available?

Auditability: Is the measure susceptible to exploitation or "gaming" that would be undetectable in an audit? Measures should not be susceptible to manipulation that would be undetectable in an audit.

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

2020 Submission:

Statistical results of construct validity testing

Table 4a. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Hospitalization for Mental Illness* Performance Scores Within Measure – **Commercial** Plans, calendar year 2018 data

Rate	Correlation Coefficient: 30-day
7-day	0.90*

*Significant at p < 0.001

Table 4b. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Hospitalization for Mental Illness* Performance Scores Within Measure – **Medicaid** Plans, calendar year 2018 data

Rate	Correlation Coefficient: 30-day
7-day	0.93*

*Significant at p < 0.001

Table 4c. Health-Plan Level Pearson Correlation Coefficients Among *Follow-Up After Hospitalization for Mental Illness* Performance Scores Within Measure – **Medicare** Plans, calendar year 2018 data

Rate	Correlation Coefficient: 30-day
7-day	0.91*

*Significant at p < 0.001

Table 5a. Results of Pearson Correlation Coefficient for Commercial, Medicaid and Medicare health plans for the *Follow-Up After Hospitalization for Mental Illness* and *Follow-Up After Emergency Department Visit for Mental Illness* Measures, Calendar Year 2018 Data

Product Line	Rate	Correlation Coefficient
Commercial	FUH 7-day and FUM 7-day	.497
*	(N=, p value =)	(316, p < 0.001)
*	FUH 30-day and FUM 30-day	.609
*	(N=, p value =)	(316, p < 0.001)
*	FUH 7-day and FUM 30-day	.555

Product Line	Rate	Correlation Coefficient
*	(N=, p value =)	(316, p < 0.001)
*	FUH 30-day and FUM 7-day	.533
*	(N=, p value =)	(316, p < 0.001)
Medicaid	FUH 7-day and FUM 7-day	.476
*	(N=, p value =)	(156, p <0.001)
*	FUH 30-day and FUM 30-day	.514
*	(N=, p value =)	(156, p <0.001)
*	FUH 7-day and FUM 30-day	.524
*	(N=, p value =)	(156, p <0.001)
*	FUH 30-day and FUM 7-day	.452
*	(N=, p value =)	(156, p <0.001)
Medicare	FUH 7-day and FUM 30-day	.537
*	(N=, p value =)	(243, p <0.001)
*	FUH 30-day and FUM 30-day	.630
*	(N=, p value =)	(243, p <0.001)
*	FUH 7-day and FUM 30-day	.585
*	(N=, p value =)	(243, p <0.001)
*	FUH 30-day and FUM 30-day	.555
*	(N=, p value =)	(243, p <0.001)

N = the number of plans reporting both indicators.

*cell intentionally left blank

2016 Update
ICD-10 CONVERSION

Summary of Stakeholder Comments Received

NCQA posted ICD-10 codes for public review and comment in March 2011 and March 2012. Comments received helped to ensure we were mapping the codes correctly.

2012 Submission

Results of face validity assessment

Step 1: The Follow-Up After Hospitalization for Mental Illness measure was developed to address a gap in care concerning follow-up care for people with mental illness. NCQA's Performance Measurement Department and the Behavioral Health MAP worked together to assess the most appropriate tools for monitoring follow-up for mental illness.

Step 2: The measure was written, field-tested, and presented to the CPM and incorporated into HEDIS in 1994.

Step 3: The measure was released for Public Comment prior to publication in HEDIS. We received and responded to comments on this measure.

Step 4: The Follow-Up After Hospitalization for Mental Illness measure was introduced in HEDIS 1994. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year.

Step 5: The Follow-Up After Hospitalization for Mental Illness measure was reevaluated in 2011/2012.

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?*)

2020 Submission:

Interpretation of construct validity testing
Correlations between individual rates within the *Follow-Up After Hospitalization for Mental Illness* measure were positive and strong (Tables 4a, 4b, 4c) across product lines. Across all product lines, correlations between the *Follow-Up After Hospitalization for Mental Illness* and the *Follow-Up After Emergency Department Visit for Mental Illness* measure rates (Table 5a) were positive and moderate. Plans with higher rates on *Follow-Up After Hospitalization for Mental Illness* tend to also have higher rates on the *Follow-Up After Emergency Department Visit for Mental Illness* measure. The results indicate that the *Follow-Up After Hospitalization for Mental Illness* measure has good validity.

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

2016 Submission:

Interpretation of systematic assessment of face validity: Our advisory panels agreed that the measures as specified will accurately differentiate quality across health plans. The measure had sufficient face validity.

2b2. EXCLUSIONS ANALYSIS

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

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

2020 Submission:

NCQA currently allows health plans to apply exclusions to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigated and validated the exclusion applied to the eligible denominator.

2016 Update: EXCLUSIONS ANALYSIS

NCQA currently allows health plans for exclusion to their results. NCQA does not collect data on exclusion for HEDIS reporting of the measure. In measure development and field testing, we investigate and validate the exclusion applied to the eligible denominator.

2012 Submission

NCQA currently allows health plans for optional exclusion to their results. NCQA does not conduct the annual analysis applied to a sample. In measure development, field testing and any re-analysis for update, we investigate and validate the effect of the reliability exclusion applied to the eligible denominator.

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*)

2020 Submission:

N/A

2016 Submission:

N/A

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

2020 Submission:

N/A

2016 Submission:

N/A

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

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

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

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

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

needed to achieve fair comparisons across measured entities.

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

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

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

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

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

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

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

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

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

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

2b3.9. Results of Risk Stratification Analysis:

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

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

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

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

2020 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 below 25th and above 75th percentile groups. The t-test method calculates a testing statistic based on the sample size, performance rate, and standard error of each plan. The test statistic is then compared against a t distribution, which is similar to a normal distribution. If the p-value of the test statistic is less than .05, then the two plans' performance is significantly different from each other.

2016 Update

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance using 2016 data. We used these two plans as examples of measured entities. However the method can be used for comparison of any two measured entities.

2012 Submission

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

Comparison of means and percentiles; analysis of variance against established benchmarks: if sample size is >400, we would use an analysis of variance.

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*)

2020 Submission:

HEDIS MY 2018 Variation in Performance across Health Plans- Commercial

Measures	N	Min	P10	P25	Mean	Median	P75	P90	Max	IQR	P value
7-day rate	361	0	0.3	0.37	0.44	0.44	0.51	0.59	0.79	0.14	p < 0.001
30-day rate	358	0	0.52	0.6	0.66	0.67	0.73	0.78	0.92	0.13	p < 0.001

HEDIS MY 2018 Variation in Performance across Health Plans- Medicare

Measures	N	Min	P10	P25	Mean	Median	P75	P90	Max	IQR	P value
7-day rate	308	0	0.13	0.18	0.28	0.25	0.34	0.46	0.68	0.16	p < 0.001
30-day rate	308	0.07	0.3	0.37	0.48	0.47	0.6	0.7	0.84	0.23	p < 0.001

HEDIS MY 2018 Variation in Performance across Health Plans- Medicaid

Measures	N	Min	P10	P25	Mean	Median	P75	P90	Max	IQR	P value
7-day rate	173	0.05	0.21	0.29	0.36	0.35	0.43	0.52	0.7	0.14	p < 0.001
30-day rate	172	0.12	0.38	0.5	0.57	0.58	0.66	0.72	0.83	0.16	p < 0.001

N = Number of plans reporting

IQR = Interquartile range

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

2016 Update

HEDIS 2016 Variation in Performance across Health Plans- Commercial

Rate	# of Plans	Avg EP	Avg.	SD	Min.	10th	25th	50th	75th	90th	IQR	P-Value
7 days	368	568	50.3%	13.1%	2.6%	34.7%	42.2%	49.8%	58.7%	65.8%	16.5%	<0.001
30 days	368	568	69.7%	11.1%	7.7%	55.4%	64.6%	70.6%	76.8%	82.5%	12.2%	<0.001

EP: Eligible Population, the average denominator size across plans submitting to HEDIS

IQR: Interquartile range

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

HEDIS 2016 Variation in Performance Across Health Plans- Medicare

Rate	# of Plans	Avg. EP	Avg.	SD	Min.	10th	25th	50th	75th	90th	IQR	P-Value
7 days	301	279	33.8%	14.9%	3.3%	15.7%	22.4%	32.0%	43.0%	55.1%	20.6%	<0.001
30 days	301	279	52.4%	17.0%	11.1%	30.6%	39.8%	53.5%	65.2%	76.2%	25.4%	<0.001

HEDIS 2016 Variation in Performance Across Health Plans- Medicaid

Rate	# of Plans	Avg. EP	Avg.	SD	Min.	10th	25th	50th	75th	90th	IQR	P-Value
7 days	166	1,182	43.6%	15.7%	0.0%	24.7%	34.2%	43.6%	55.2%	64.2%	21.0%	<0.001
30 days	168	1,169	61.2%	16.0%	8.1%	41.3%	54.1%	63.7%	72.6%	78.5%	18.5%	<0.001

Figure 1a. *Follow-Up After Hospitalization for Mental Illness -7-Day Rate*: Commercial Plans 2014-2016

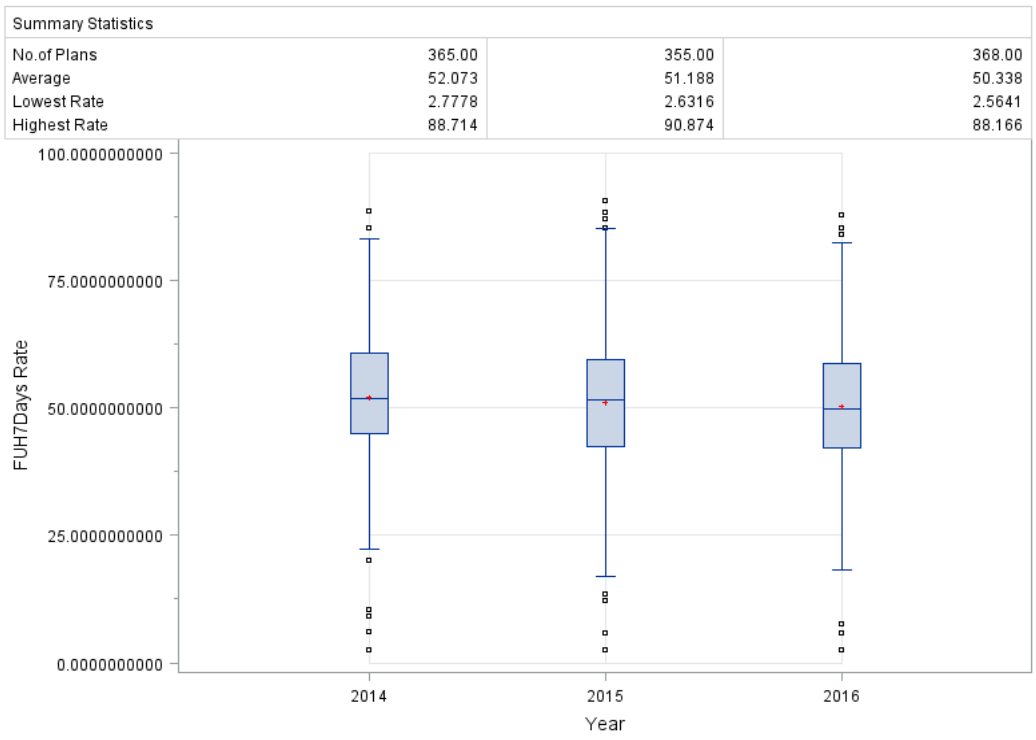


Figure 1b. ***Follow-Up After Hospitalization for Mental Illness -30-Day Rate*****: Commercial Plans 2014-2016**

Boxplot Graph for Commercial FUH 30Day Rate from 2014-2016

Summary Statistics			
No.of Plans	365.00	355.00	368.00
Average	70.790	70.061	69.696
Lowest Rate	13.636	17.073	7.6923
Highest Rate	95.313	95.522	93.343

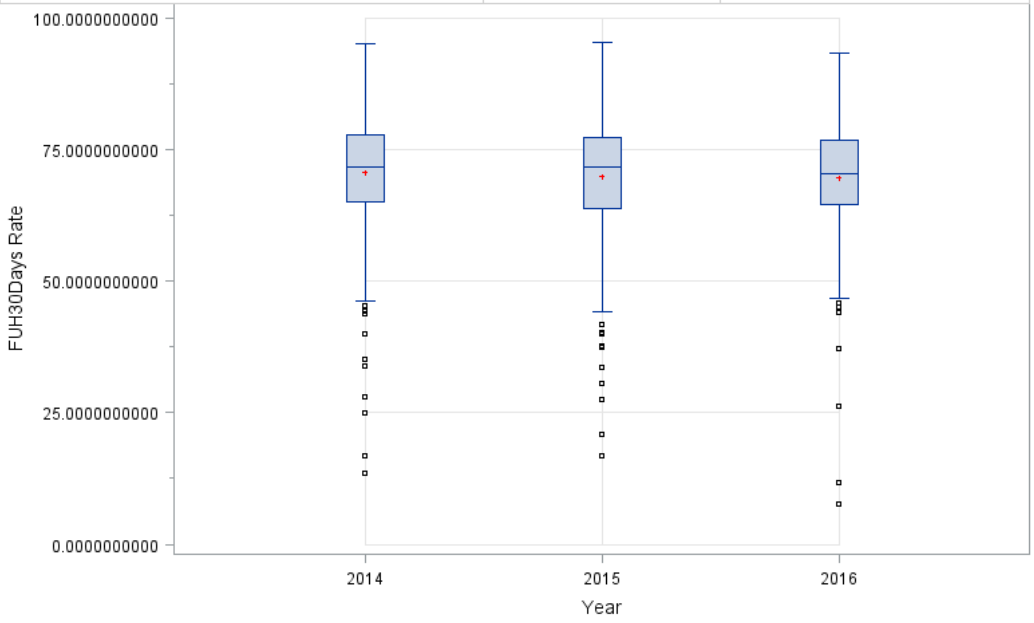


Figure 3a. ***Follow-Up After Hospitalization for Mental Illness -7-Day Rate*****: Medicare Plans 2014-2016**

Boxplot Graph for Medicare FUH 7Day Rate from 2014-2016

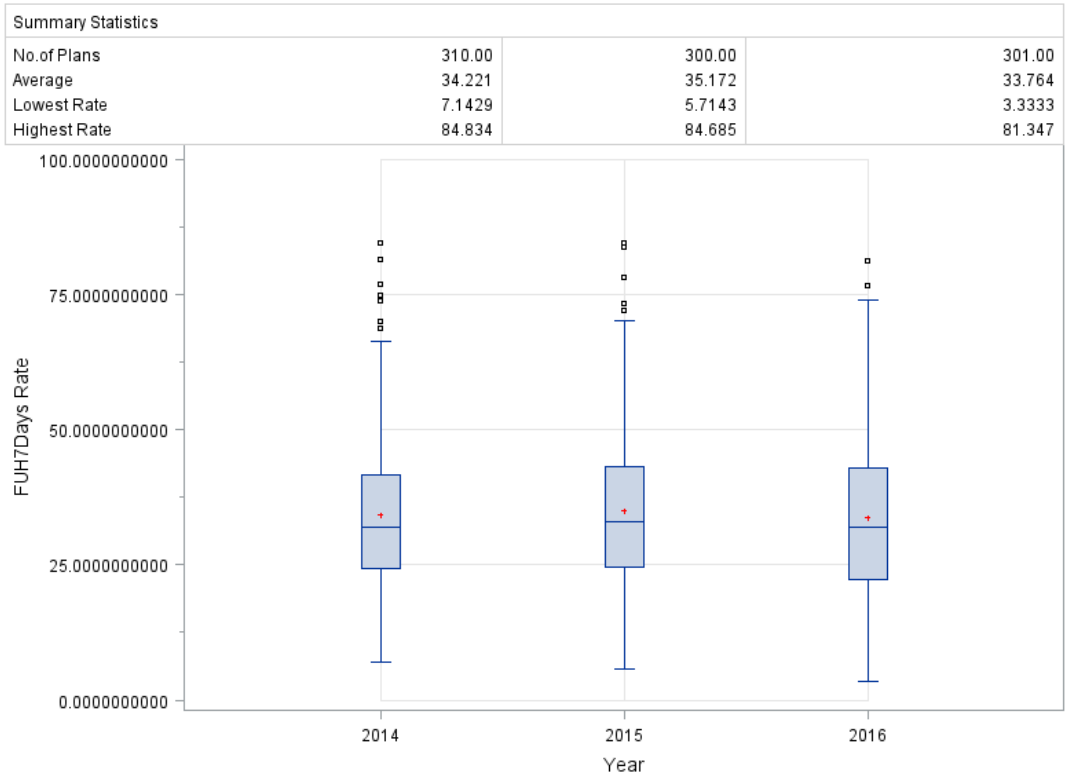


Figure 3b. ***Follow-Up After Hospitalization for Mental Illness -30-Day Rate*****: Medicare Plans 2014-2016**

Boxplot Graph for Medicare FUH 30Day Rate from 2014-2016

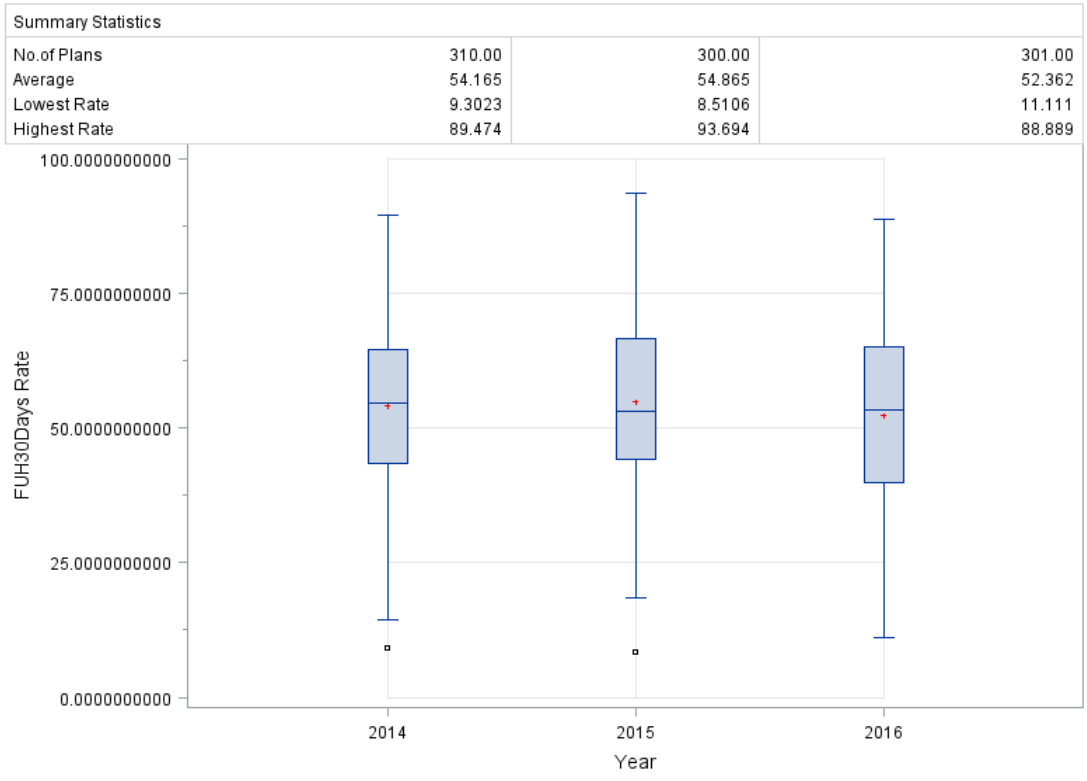


Figure 2a. ***Follow-Up After Hospitalization for Mental Illness -7-Day Rate*****: Medicaid Plans 2014-2016**

Boxplot Graph for Medicaid FUH 7Day Rate from 2014-2016

Summary Statistics			
No.of Plans	123.00	135.00	166.00
Average	41.992	43.823	43.624
Lowest Rate	1.5625	1.9231	0.0000
Highest Rate	76.757	76.256	75.000

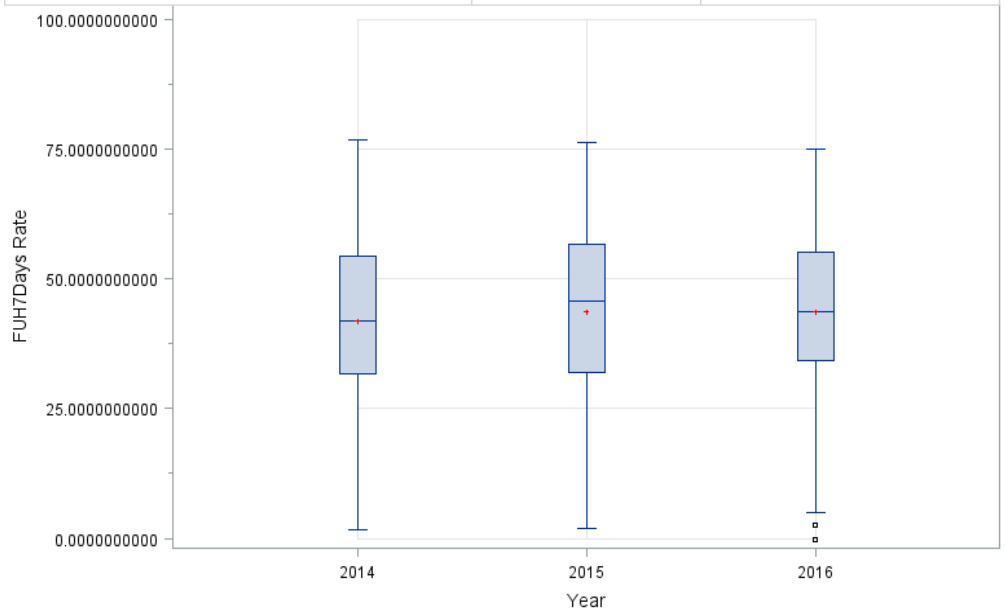
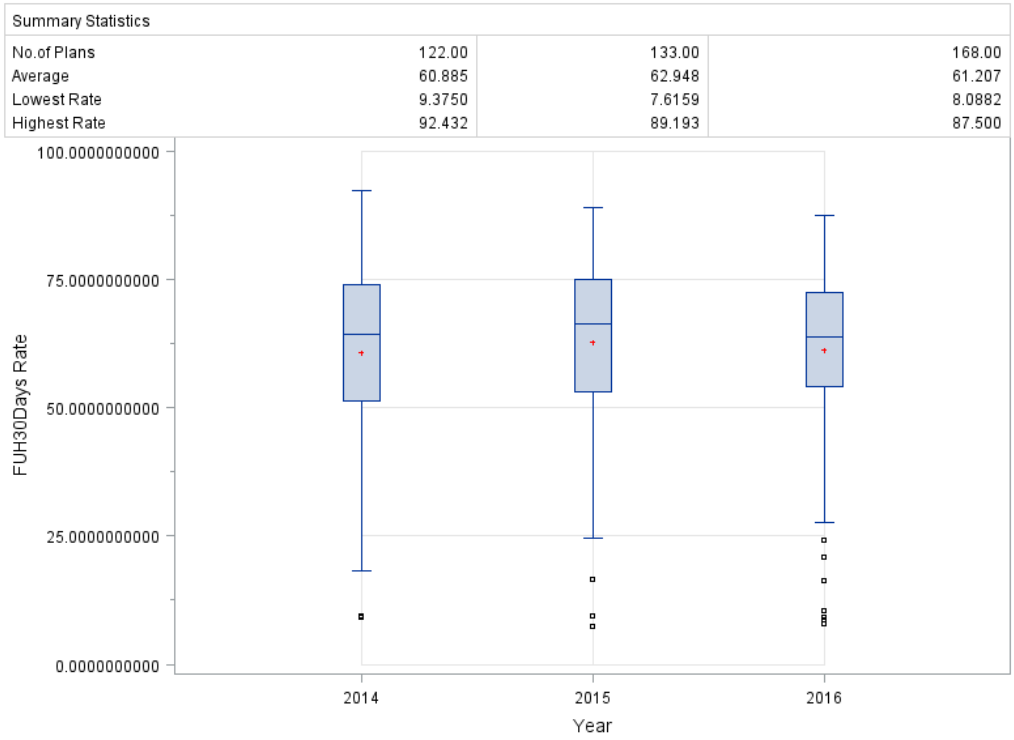


Figure 2b. ***Follow-Up After Hospitalization for Mental Illness -30-Day Rate*****: Medicaid Plans 2014-2016**

Boxplot Graph for Medicaid FUH 30Day Rate from 2014-2016



2012 Submission

7-Day Rate

Commercial

Measurement Year: 2009; 2010; 2011

N:	397	391	363
Min:	5.8	3.75	3.13
Max:	97.62	90.18	93.33
Mean:	54.01	55.96	57.22
SD:	13.1	13.75	12.88
P10:	37.93	39.22	42.05
P25:	45.26	46.54	48.74
P50:	53.85	56.01	57.04
P75:	62.96	65.19	66.13
P90:	71.23	72.76	72.07

Medicaid

Measurement Year: 2009; 2010; 2011

N:	62	71	85
Min:	2.6	8.2	10.87
Max:	78.57	87.9	86.85
Mean:	42.62	42.89	44.56
SD:	18.29	18.6	16.45

P10: 15.52 18.22 23.02
P25: 31.65 29.59 33.1
P50: 44.53 43.52 45.11
P75: 56.63 59.1 53.91
P90: 64.15 64.25 68.31

Medicare
Measurement Year: 2009; 2010; 2011
N: 193 231 257
Min: 4.23 2.13 1.67
Max: 86.67 84.21 84
Mean: 37.97 38 37.8
SD: 17.55 18.33 18.02
P10: 15.57 13.7 15.38
P25: 23.26 23.86 24.24
P50: 36.88 36.84 37.44
P75: 51.39 50 48.45
P90: 60.32 63.49 63.93

30-Day Rate
Commercial
Measurement Year: 2009; 2010; 2011
N: 397 391 364
Min: 21.74 21.21 13.58
Max: 98.61 97.32 100
Mean: 74.1 74.68 75.93
SD: 10.31 10.8 10.49
P10: 60 61.57 64.89
P25: 67.94 68.82 71.02
P50: 74.74 76 76.38
P75: 81.82 82.21 82.43
P90: 85.96 86.29 87.2

Medicaid
Measurement Year: 2009; 2010; 2011
N: 61 70 82
Min: 18.07 15.63 22.7
Max: 87.5 91.67 87.79
Mean: 61.67 60.22 63.83
SD: 18.25 19.14 16.19
P10: 37.27 31.79 36
P25: 49.6 49.02 57.14
P50: 64.29 62.63 66.6

P75:	75.65	74.28	74.62
P90:	81.23	83.57	82.56

Medicare

Measurement Year: 2009; 2010; 2011

N:	193	230	254
Min:	9.86	5.77	5.95
Max:	100	96.15	93.33
Mean:	56.32	55.99	56.69
SD:	18.38	19.17	18.73
P10:	30	27.3	29.79
P25:	43.82	42.11	44.87
P50:	58.1	58.23	57.95
P75:	71.43	71.88	70
P90:	78.18	79.72	80

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?*)

2020 Submission

For the 7-day rate, there is a 14 percentage point gap in performance between Commercial plans at the 25th and 75th percentiles, a 16 percentage point gap for Medicare plans, and a 14 percentage point gap for Medicaid plans. For the 30-day rate, there is a 13 percentage point gap in performance between Commercial plans at the 25th and 75th percentiles, a 23 percentage point gap for Medicare plans, and a 16 percentage point gap for Medicare plans. The difference in performance between plans in the 25th percentile and 75th percentile is statistically significant for both rates across all product lines.

2016 Submission:

The results above indicate there is a 12-25% gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the Medicare health plans which show a 20.6-25.4 percentage point gap between 25th and 75th percentile plans.

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

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

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

factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

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

2020 Submission

This measure has only one set of specifications.

2016 Submission:

N/A

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*)

2020 Submission

This measure has only one set of specifications.

2016 Submission:

N/A

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*)

2020 Submission

This measure has only one set of specifications.

2016 Submission:

N/A

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

2020 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

2016 Submission:

Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

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*)

2020 Submission: 2020 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 feasibility of the measure when widely implemented in the field. This analysis includes an assessment of how many plans report valid rates vs. rates that are materially biased (or have other issues, such as small denominators). These considerations are weighed in the deliberation process before measures are approved for public reporting.

2016 Submission:

Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

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 non-responders) 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*)

2020 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.

2016 Submission:

Plans collect this measure using all administrative data sources. NCQA's audit process checks that plans' measure calculations are not biased due to missing data.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in a combination of electronic sources

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

N/A

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

Attachment:

3c. Data Collection Strategy

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

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

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

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

The HEDIS Compliance Audit addresses the following functions:

1) information practices and control procedures
2) sampling methods and procedures
3) data integrity
4) compliance with HEDIS specifications
5) analytic file production
6) reporting and documentation

In addition to the HEDIS Audit, NCQA provides a system to allow “real-time” feedback from measure users. Our Policy Clarification Support System receives thousands of inquiries each year on over 100 measures. Through this system NCQA responds to questions in order to prevent possible errors or inconsistencies in the implementation of the measure. Input from NCQA auditing and the Policy Clarification Support System informs the annual updating of all HEDIS measures including updating value sets and clarifying the specifications. Measures are re-evaluated on a periodic basis and when there is a significant change in evidence. During re-evaluation information from NCQA auditing and Policy Clarification Support System is used to inform evaluation of the usability and feasibility of the measure.

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

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
Quality Improvement (Internal to the specific organization)	Public Reporting Health Plan Ranking https://healthinsuranceratings.ncqa.org/ Medicaid Child Core Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/child-core-set/index.html Medicare Adult Core Set https://www.medicaid.gov/medicaid/quality-of-care/performance-measurement/adult-core-set/index.html Hospital Compare https://www.medicare.gov/hospitalcompare/search.html? Inpatient Psychiatric Facility Quality Reporting https://www.qualitynet.org/ipf/ipfqr Qualified Health Plan (QHP) Quality Rating System (QRS) https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/QRS-and-QHP-Enrollee-Survey-Technical-Guidance-for-2020-508.pdf Payment Program Quality Payment Program https://qpp.cms.gov/mips/explore-measures Regulatory and Accreditation Programs Health Plan Accreditation https://www.ncqa.org/programs/health-plans/health-plan-accreditation-hpa/ Quality Improvement (external benchmarking to organizations) Quality Compass https://www.ncqa.org/programs/data-and-information-technology/data-purchase-and-licensing/quality-compass/ Annual State of Health Care Quality http://www.ncqa.org/tabid/836/Default.aspx

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

HEALTH PLAN RATINGS/REPORT CARDS: This measure is used to calculate health plan ratings which are reported on the NCQA website. These ratings are based on performance on HEDIS measures among other factors. In 2019, a total of 255 Medicare health plans, 515 commercial health plans and 188 Medicaid health plans across 50 states were included in the ratings.

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 2019, the report included results from calendar year 2018 for health plans covering a record 136 million people, or 43 percent of the U.S. population.

MEDICAID CHILD CORE SET: This measure is included in the Medicaid Child Core Set which is a set of children's health care quality measures developed as part of the Children's Health Insurance Program (CHIP) Reauthorization Act for voluntary use by State Medicaid and CHIP programs. The data collected with these measures will help CMS to better understand the quality of health care children receive through Medicaid and

CHIP and assist CMS and states in moving toward a national system for quality measurement, reporting, and improvement. As per the CHIPRA legislation, state data derived from the core measures will become part of the Secretary's annual report on the quality of care for children in Medicaid and CHIP. The Secretary's annual report summarizes state-specific and national measurement information on the quality of health care furnished to children enrolled in Medicaid and CHIP.
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, as of 2014, state data on the adult quality measures is part of the Secretary's annual report on the quality of care for adults enrolled in Medicaid.
HEALTH PLAN ACCREDITATION: This measure is used in scoring for accreditation of Medicare Advantage Health Plans. In 2019, 336 commercial health plans covering 87 million lives and 77 Medicaid health plans covering 9.1 million lives were accredited. Health plans are scored based on performance compared to benchmarks
QUALITY COMPASS: This measure is used in Quality Compass which is an indispensable tool used for selecting a health plan, conducting competitor analysis, examining quality improvement and benchmarking plan performance. Provided in this tool is the ability to generate custom reports by selecting plans, measures, and benchmarks (averages and percentiles) for up to three trended years. Results in table and graph formats offer simple comparison of plans' performance against competitors or benchmarks.
HOSPITAL COMPARE: This measure is used in Hospital Compare which helps improve quality of care by sharing objective, easy to understand data on hospital performance as well as consumer perspectives.
INPATIENT PSYCHIATRIC FACILITY QUALITY REPORTING: This measure is used in the Inpatient Psychiatric Facility Quality Reporting program which provides consumers with quality of care information to make informed decisions about their healthcare options. This program is intended to encourage clinicians and psychiatric facilities to the quality if inpatient care via awareness and reporting of best practices for respective facilities and types of care.
QUALIFIED HEALTH PLAN (QHP) QUALITY RATING SYSTEM (QRS): This measure is used in the Qualified Health Plan (QHP) Quality Rating System (QRS) which provides comparable information to consumers about the quality of health care services and QHP enrollee experience offered in the Marketplaces.
QUALITY PAYMENT PROGRAM:
The Quality Payment Program (QPP) is a quality and cost incentive program that uses payment adjustments to promote high quality and high value care delivery by eligible clinicians (EC). QPP provides performance-based payment adjustments to ECs, both negative and positive, for services furnished to Medicare Part B beneficiaries. EC performance is graded on quality measure performance, cost of care, engagement in clinical practice improvement activities, and use of Certified EHR Technology (CEHRT). Performance can be reported at the individual (clinician) or group (practice) level. In 2017, 1,006,319 ECs participated in MIPS, representing 95% of all eligible clinicians across the 50 states. 54% participated as a part of a group, 12% as individual clinicians, and 34% as a part of an Advanced Payment Model.

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 3c1.

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

Describe how feedback was obtained.

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

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

In general, health plans have considered this measure feasible for reporting using the administrative data collection method. Questions received were about clarification of the specifications, such as confirmation that a type of provider met the definition of mental health providers and research supporting the measure. NCQA responded to all questions to ensure consistent implementation of the measure.

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

This measure has been deemed a priority measure by NCQA and other entities, as illustrated by its use in programs such as the Medicaid Child and Adult Core Sets, CMS EHR Incentive Program and CMS Physician Quality Reporting Initiative.

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 specification to include clarifying text and additional examples to further support determining numerator compliance.

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.

2017 to 2018 data shows relatively stable performance and room for improvement across Commercial and Medicaid plans. The mean performance for the 7-day rate was .46 in 2017 and .44 in 2018 among Commercial plans. Among Medicaid Plans, the mean performance for the 7-day rate was .37 in 2017 and .36 in 2018. Performance rates for the 30-day rate also remained relatively stable from 2017 to 2018 for commercial and Medicaid plans. Medicare performance rates declined slightly across both rates; in 2017, the mean 7-day performance rate was .32, declining to .28 in 2018. The 30-day mean performance rate declined from .53 in 2017 to .48 in 2018. Across all product lines, there continues to be fairly large variation between the 10th and 90th percentiles, suggesting room for improvement. For example, among Medicare plans, the 2018 7-day rate ranged from 13% for plans in the 10th percentile to 46% among plans in the 90th percentile.

4b2. Unintended Consequences

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

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

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

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

There were no identified 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.

No

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

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

No

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

N/A

5b. Competing Measures

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

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

N/A

Appendix

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

No appendix **Attachment:**

Contact Information

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

Co.2 Point of Contact: Bob, Rehm, nqf@ncqa.org, 202-955-1728-

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

Co.4 Point of Contact: Brittany, Wade, wade@ncqa.org, 202-530-0463-

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.

NCQA 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

Connie Horgan, ScD, Brandeis University

Laura Jacobus-Kantor, PhD, SAMSA, HHS

Jeffry Meyerhof, MD, Optum Behavioral Health

Harold Pincus, MD, Columbia University

Michael Schoenbaum, PhD, National Institute of Mental Health

John H. Straus, MD, Beacon Health Options

NCQA TECHNICAL MEASUREMENT ADVISORY PANEL

Andy Amster, MSPH, Kaiser Permanente

Jennifer Brudnicki, MBA, Inovalon Inc.

Lindsay Cogan, PhD, MS, New York State Department of Health

Kathryn Coltin, MPH, Independent Consultant

Mike Farina, RPh, MBA, Capital District Physicians' Health Plan

Marissa Finn, MBA, CIGNA

Scott Fox, MS, MEd, FAMIA, The MITRE Corporation

Carlos Hernandez, CenCal Health

Harmon Jordan, ScD, Westat

Virginia Raney, LCSW, Center for Medicaid and CHIP Services

Lynne Rothney-Kozlak, MPH, Rothney-KozlakConsulting, LLC

Laurie Spoll, Aetna

Geriatric Measurement Advisory Panel (GMAP):

Wade M. Aubry, MD, University of California—San Francisco

Arlene S. Bierman, MD, MS, Agency for Healthcare Research and Quality

Patricia A. Bomba, MD, MACP, FRCP, Excellus BlueCross BlueShield

Nicole Brandt, PharmD, MBA, BCGP, BCPP FASCP, University of Maryland, School of Pharmacy

Jennie Chin Hansen, RN, MS, FAAN, Geriatric Expert

Joyce Dubow, MUP, Consumer Representative

Pete Hollmann, MD, Brown Medicine

Jeff Kelman, MD, Department of Health and Human Services

Karen J. Nichols, MD, Trinity-Health PACE

Steven L. Philips, MD, CMD, Geriatric Specialty Care

Erwin Tan, MD, American Association of Retired Persons

Eric G. Tangalos, MD, Mayo Clinic

Dirk Wales, MD, PsyD, Axial Healthcare

Joan Weiss, PhD, RN, CRNP, FAAN, Health Resources and Services Administration

Neil Wenger, MD, University of California, Los Angeles

COMMITTEE ON PERFORMANCE MEASUREMENT

Andrew Baskin, MD CVS Health/Aetna

Elizabeth Drye, MD, SM Yale School of Medicine

Andrea Gelzer, MD, MS, FACP AmeriHealth Caritas

Kate Goodrich, MD, MHS Centers for Medicare & Medicaid Services

David Grossman, MD, MPH Washington Permanente Medical Group

Christine S. Hunter, MD (Co-Chair) Independent Board Director

David K. Kelley, MD, MPA Pennsylvania Department of Human Services

Jeff Kelman, MD, MMSc Department of Health and Human Services

Nancy Lane, PhD Independent Consultant

Bernadette Loftus, MD Independent Consultant

Adrienne Mims, MD, MPH, AGSF, FAAFP Alliant Health Solutions

Amanda Parsons, MD, MBA MetroPlus

Wayne Rawlins, MD, MBA ConnectiCare

Misty Roberts, MSN, RN, CPHQ, PMP Humana

Rodolfo Saenz, MD, MMM, FACOG Riverside Medical Clinic

Marcus Thygeson, MD, MPH (Co-Chair) Bind Benefits

JoAnn Volk, MA Georgetown University Liaisons

Rose Baez, RN, MSN, MBA, CPHQ Blue Cross Blue Shield Association

Jeff Brady, MD, MPH Agency for Healthcare Research and Quality

Ron Kline, MD Office of Personnel Management

Elisa Munthali, MPH National Quality Forum

Chinwe Nwosu, MS America's Health Insurance Plans

Chesley Richards, MD, MPH, FACP Centers for Disease Control and Prevention

Anecia Suneja, CNS-BC Veteran's Health Administration

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 1994

Ad.3 Month and Year of most recent revision: 07, 2020

Ad.4 What is your frequency for review/update of this measure? Approximately every 3-5 years, sooner if the clinical guidelines have changed significantly.

Ad.5 When is the next scheduled review/update for this measure? 12, 2021

Ad.6 Copyright statement: © 2020 by the National Committee for Quality Assurance

1100 13th Street, NW, Third Floor

Washington, DC 20005

Ad.7 Disclaimers: These performance Measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications.

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

Ad.8 Additional Information/Comments: NCQA Notice of Use. Broad public use and dissemination of these measures is encouraged and NCQA has agreed with NQF that noncommercial uses do not require the consent of the measure developer. Use by health care physicians in connection with their own practices is not commercial use. Commercial use of a measure requires the prior written consent of NCQA. As used herein, "commercial use" refers to any sale, license or distribution of a measure for commercial gain, or incorporation of a measure into any product or service that is sold, licensed or distributed for commercial gain, even if there is no actual charge for inclusion of the measure.

These performance measures were developed and are owned by NCQA. They are not clinical guidelines and do not establish a standard of medical care. NCQA makes no representations, warranties or endorsement about the quality of any organization or physician that uses or reports performance measures, and NCQA has no liability to anyone who relies on such measures. NCQA holds a copyright in these measures and can rescind or alter these measures at any time. Users of the measures shall not have the right to alter, enhance or otherwise