NATIONAL QUALITY FORUM
Driving measurable health improvements together

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 0712

Measure Title: Depression Assessment with PHQ-9/ PHQ-9M

Measure Steward: Minnesota Community Measurement

Brief Description of Measure: The percentage of adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) with a diagnosis of major depression or dysthymia who have a completed PHQ-9 or PHQ-9M tool during a four month measurement period.

Developer Rationale: Adults:

Depression is a common and treatable mental disorder. The Centers for Disease Control and Prevention states that an estimated 6.6% of the U.S. adult population (14.8 million people) experiences major depressive disorder during any given 12-month period. Additionally, dysthymia accounts for an additional 3.3 million Americans. In 2006 and 2008, an estimated 9.1% of U.S. adults reported symptoms for current depression.1 Persons with a current diagnosis of depression and a lifetime diagnosis of depression or anxiety were significantly more likely than persons without these conditions to have cardiovascular disease, diabetes, asthma and obesity and to be a current smoker, to be physically inactive and to drink heavily.2 People who suffer from depression have lower incomes, lower educational attainment and fewer days working days each year, leading to seven fewer weeks of work per year, a loss of 20% in potential income and a lifetime loss for each family who has a depressed family member of $300,000.3 The cost of depression (lost productivity and increased medical expense) in the United States is $83 billion each year.4
Prevalence updates: 2019 National Survey on Drug Use and Health (NSDUH) estimates that 19.4 million adults or 7.8% had at least one major depressive episode with the highest prevalence of 15.2% among individuals aged 18 - 25.14
Adolescents and Adults:
The Centers for Disease Control and Prevention states that during 2009-2012 an estimated 7.6% of the U.S. population aged 12 and over had depression, including 3% of Americans with severe depressive symptoms. Almost 43% of persons with severe depressive symptoms reported serious difficulties in work, home and social activities, yet only 35% reported having contact with a mental health professional in the past year.5
Depression is associated with higher mortality rates in all age groups. People who are depressed are 30 times more likely to take their own lives than people who are not depressed and five times more likely to abuse

drugs.6 Depression is the leading cause of medical disability for people aged 14 – 44.7 Depressed people lose 5.6 hours of productive work every week when they are depressed, fifty percent of which is due to absenteeism and short-term disability.
Adolescents:
In 2014, an estimated 2.8 million adolescents age 12 to 17 in the United States had at least one major depressive episode in the past year. This represented 11.4% of the U.S. population. The same survey found that only 41.2 percent of those who had a Major Depressive Episode received treatment in the past year.8 The 2013 Youth Risk Behavior Survey of students grades 9 to 12 indicated that during the past 12 months 39.1% (F) and 20.8% (M) indicated feeling sad or hopeless almost every day for at least 2 weeks, planned suicide attempt 16.9% (F) and 10.3% (M), with attempted suicide 10.6% (F) and 5.4% (M).9 Adolescent-onset depression is associated with chronic depression in adulthood.10 Many mental health conditions (anxiety, bipolar, depression, eating disorders, and substance abuse) are evident by age 14.11 The 12-month prevalence of MDEs increased from 8.7% in 2005 to 11.3% in 2014 in adolescents and from 8.8% to 9.6% in young adults (both $P < .001$). The increase was larger and statistically significant only in the age range of 12 to 20 years. The trends remained significant after adjustment for substance use disorders and sociodemographic factors. Mental health care contacts overall did not change over time; however, the use of specialty mental health providers increased in adolescents and young adults, and the use of prescription medications and inpatient hospitalizations increased in adolescents. 12 In 2015, 9.7% of adolescents in MN who were screened for depression or other mental health conditions, screened positively.13

Numerator Statement: Adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) included in the denominator who have at least one PHQ-9 or PHQ-9M tool administered and completed during a four month measurement period.

Denominator Statement: Adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) with a diagnosis of major depression or dysthymia.

Denominator Exclusions: Patients who die, are a permanent resident of a nursing home or are enrolled in hospice are excluded from this measure. Additionally, patients who have a diagnosis of bipolar or personality disorder, schizophrenia or psychotic disorder, or pervasive developmental disorder are excluded.

Measure Type: Process

Data Source: Electronic Health Records

Level of Analysis: Clinician: Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: 01/17/2011

Most Recent Endorsement Date: 03/06/2015

Preliminary Analysis: Maintenance of Endorsement

To maintain NQF endorsement, endorsed measures are evaluated periodically to ensure that the measure 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 are 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 description for this measure:

- This is maintenance process measure at the clinician group /practice level that assesses the proportion of adolescent patient (12 to 17 years of age) and adult patients (18 years of age or older) who have had at least one PHQ-9 or PHQ-9M tool administered during a four-month measurement period.
- The developer provides a logic model that depicts the assessment of the diagnosis of major depressive disorder (MDD) or dysthymia using the PHQ-9/PHQ-9M PROM, which leads to treatment with medication and/or therapy, where progress can be assessed with the PHQ-9/PHQ-9 PROM.

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 2015:

- This measure is a paired process measure that seeks to promote frequent use of the PHQ-9 and supports the two additional Minnesota (MN) Community Measurement outcome measures submitted (#0710 and #0711). This measure, unlike the outcome measures, examines the entire population that has depression or dysthymia, regardless of the PHQ-9 score.
- During the prior review, there was general agreement that depression and dysthymia are common illnesses occurring in nine percent of the population and the measure was supported by evidence.

Changes 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:

- The developer does not provide direct evidence of measurement of the PHQ-9 in isolation, as compared to non-performance of the PHQ-9, in linking to improved outcomes. Rather the evidence provided is about PHQ-9 being a validated tool and the use of the PHQ-9 being part of collaborative care, which PHQ-9 performance is one component.
- The developer cites the Institute for Clinical Systems Improvement (ICSI): Adult Depression in Primary Care Guideline as support for the adult portion of this measure.
    - The guidelines are based on an ICSI systematic review of several studies and random control trials (RCTs): four RCTs (high evidence grade), one observational study (Low evidence grade) and two cohort studies (low evidence grade).
    - The guideline states that PHQ-9 has been a validated tool for measuring depression severity and is an effective management tool for routine use in subsequent visits.
    - The developer highlights two main recommendations from the guidelines:
        - Comprehensive Treatment Plan with Shared Decision-Making Collaborative Care Model: A collaborative care approach is recommended for patients with depression in primary care (Quality of Evidence: High; Strength of Recommendation: Strong)

        - Establish Follow-Up Plan: Clinicians should establish and maintain follow-up with patients (Quality of Evidence: Low Strength of Recommendation: Strong)
- The developer also cited the Guidelines for Adolescent Depression in Primary Care (GLAD-PC): Part I. Practice Preparation, Identification, Assessment, and Initial Management for the adolescent population.
    - Grading of evidence is based in a 1–5 system, with 1 to 5 corresponding to strongest to weakest evidence. Strength of recommendation is broken down into 4 categories: very strong (>90% agreement), strong (>70% agreement), fair (>50% agreement), and weak (<50% agreement).
    - The developer highlights the following recommendation:
        - Identification and Surveillance Recommendation 2: Patients with depression risk factors (e.g., a history of previous depressive episodes, a family history, other psychiatric disorders, substance use, trauma, psychosocial adversity, frequent somatic complaints, previous high-scoring screens without a depression diagnosis, etc.) should be identified (grade of evidence: 2; strength of recommendation: very strong) and systematically monitored over time for the development of a depressive disorder by using a formal depression instrument or tool (targeted screening) (grade of evidence: 2; strength of recommendation: very strong).
        - Treatment Recommendation 1: PC clinicians should work with administration to organize their clinical settings to reflect best practices in integrated and/or collaborative care models (e.g., facilitating contact with psychiatrists, case managers, embedded therapists). (grade of evidence: 4; strength of recommendation: very strong).
        - Ongoing Management Recommendation 1: Systematic and regular tracking of goals and outcomes from treatment should be performed, including assessment of depressive symptoms, and functioning in several key domains. These include home, school, and peer settings (grade of evidence: 4; strength of recommendation: very strong).

Question for the Committee:

- *The developer does not provide data that directly links the performance of the PHQ-9 with outcomes. Rather, they show that collaborative care, an element of which is to perform the PHQ-9 has strong evidence that is linked to outcomes.*
- *Does the Standing Committee agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Guidance from the Evidence Algorithm

Box 1 (No, this is a process measure) -> Box 3 (No, no empirical evidence was submitted link PHQ-9 performance in isolation to outcomes) -> Box 10 (No) -> Box 11 (Yes, if we count the empirical evidence submitted for PHQ-9 as a component of collaborative care) -> Box 12 (Insufficient evidence needs to be reviewed by the Committee).

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

Rationale: The developer does not show that performing the PHQ-9 in of itself is associated with any improvements in outcomes, compared to not performing this screening, especially within the four-month timeframe.

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

Maintenance measures – increased emphasis on gap and variation

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

- In MN statewide reporting for year 2020, the average was 77.7 percent for adults (n = 248,163) and 78.4 percent for adolescents (n = 19,574). There was variability among medical groups, which is displayed by the range of results (25 to 100% and 8 to 100% respectively). Standard deviation and interquartile range were not reported.

Disparities

- The developer does not present disparity data on the current measure, but rather on the related outcome measures (NQF #0710e, NQF #0711, NQF #1884, and NQF #1885).
    - From outcome measure data, the developer notes that adults who are Black, Indigenous/Native, Multi-Race or Hispanic/Latinx are among those who have significantly lower rates of depression follow-up, response and remission at six months compared to the race/ethnicity averages. Adults who are Asian have significantly lower rates of depression response and remission at six months.
    - Also from outcome measure data, adolescents who are Black have significantly lower rates of follow-up, response and remission at six months compared to the race/ethnicity averages. Adolescents who are Indigenous/Native have significantly lower rates of follow-up at six months compared to the race average.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Performance gap data provided is from the state of Minnesota. Is this data generalizable on a national level?*
- *Do you have concerns with the lack of disparities data presented specific to this measure?*

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

Committee Pre-evaluation Comments:

1a. Evidence

- This is a process measure for two different age groups (12-17) and adults. It assesses % of patients that had at least one PHQ 9 or PHQ-9M any time during a calendar 4 month interval that likely aligns with quarterly reports that this company generates for their clients. Thus the screening could occur at variable time points within a patient's episode of care. It appears that the data are aggregated at the medical group and clinic level so their customers can roughly compare performance across medical groups or clinics. However the scientific evidence provided are summaries of clinical guidelines for depression and how PHQ 9, a validated depression screener, is commonly used in collaborative care models (most evidence for adults). I agree with preliminary rating of insufficient, but more concerned about interpretation of data from this and the other MN measures. it's difficult to pinpoint a target for QI for this and the outcome (response, remission) MN measures.
- This measure would be stronger if the developer linked PHQ9 data to improved outcomes (e.g. lower hospitalization/ED visits, increased score on PHQ9)

- Concerns that the developer does not show that performing the PHQ-9 in of itself is associated with any improvements in outcomes, compared to not performing this screening.
- Maintenance Process measure-Yes new evidence has been provided however the evidence is insufficient in determining PHQ-9 performance in isolation to outcomes
- While first step in pathway to measurement based care, I believe it is time to move beyond measurement (and we have a multitude of measures that look at outcomes,not simply measurement)
- I am surprised the developer could not locate empirical evidence that administering the PHQ9 improves outcomes, but that could also be because the PHQ9 is the most commonly used measure to detect depression, so not administering it essentially means missing the diagnoses and therefore lacking comparison. If clinical guidelines unanimously support its use, then I think that is sufficient. Without administering the PHQ9, diagnoses will be missed, as will opportunities for treatment.
- The logic model as MNCM describes it relies on the fact that Collaborative Care Management is the by far the most effective model in terms of response/remission rate for depression and using the PHQ-9 is an integral part of it. Using a valid and reliable tool to measure intensity of symptom (such as the PHQ-9) IS a necessary element to measure response and remission rates (and is useful to inform progress and symptoms that require more focus and intervention.) As a stand-alone measure it does not directly correlate with outcomes. If a system/provider is trying to improve ones outcomes it informs them about the potential causes for poor outcomes i.e. how much of the result is secondary to unreliable measurement.

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

- [Standing Committee feedback]

Criteria 2: Scientific Acceptability of Measure Properties

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

Evaluators: Staff

2a. Reliability: Specifications and Testing

For maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures.

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

For maintenance measures – less emphasis if no new testing data provided.

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Specifications:

- Measure specifications are clear and precise.
- There are no concerns with the measure specifications
- Specifications have been updated since last endorsement. Changes include:
  - Incorporating adolescents ages 12-17 into the measure
  - Adding PHQ-9M PRO tool to the measure (modified for teens)
  - Modification of the exclusion value set of personality disorder
  - Addition of exclusions for schizophrenia and pervasive developmental disorder

- It is noted by the developer that the PHQ-9/PHQ-9 PROM is validated for both the assessment and diagnosis of depression and for monitoring continuing outcomes of treatment.

Reliability Testing:

- Due to the condition's chronic episodic nature, no sampling is allowed and the full population of eligible patients, regardless of payer, is included.
    - Sites represent all primary care and behavioral health (psychiatry) clinics in Minnesota and bordering cities in other states that wish to participate. Clinics represent urban and rural, large multi-specialty health care systems, medium and small practices that care for adult patients with depression. 103 medical groups representing 615 clinics were included in the testing of this measure, representing 227,127 adults and 12,616 adolescents.
    - Testing used adult patients with dates of service 10/1/2019 to 1/31/2020 reported in 2020, and adolescent patients age 12 to 17 with dates of service 1/1/2019 to 12/31/2020.
    - Clinics must have greater than or equal to 30 patients in the denominator to be included.
- Reliability testing conducted at the Accountable Entity Level:
    - Empirical testing of computed performance scores for reportable clinics was conducted using a beta binomial model (signal-to-noise).
        - The developer found that using 601 clinics and 227,000 patients, the PHQ-9 Assessment- Adults had an average reliability score of 0.932903. Interquartile (IQR) range was not provided.
        - The developer found that using 142 clinics and 12,616 patients, the PHQ-9 Assessment- Adolescents had an average reliability score of 0.878959. IQR was not provided.
        - The developer states that a beta-binomial reliability score of greater than 0.70 indicates the ability to distinguish higher performing clinics from lower performing clinics.

- Reliability of the PHQ-9
    - Reliability of the PROM has been validated in the literature in adult populations, including calculating a Cronbach's alpha and test-retest reliability.
    - The PHQ-9 has also been validated in adolescent populations (ages 13 to 17), but the PHQ-9M Modified for Teens has not undergone separate validation studies. The developer attests that this version has essentially the same nine questions as the PHQ-9 with slight wording variation for an adolescent population.
        - Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study
        - Cronbach's alpha of 0.86 in the PHQ OBGYN Study
        - Test-retest showed the correlation between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours was 0.84, and the mean scores were nearly identical (5.08 vs 5.03).
        - The developer states that this testing demonstrates the PHQ-9 tool is appropriate for measuring patient outcomes related to depression.

Questions for the Committee regarding reliability:

- *Do you have any concerns that the measure cannot be consistently implemented (i.e., are measure specifications adequate)?*
- *Do you have any concerns that the PHQ-9M has not undergone reliability testing?*

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

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

For maintenance measures – less emphasis if no new testing data provided.

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.

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

Validity Testing

- Validity of the PROM- PHQ-9:
    - The developer references testing of construct validity in the literature, using mental health professional re-interview as the criterion standard
        - Sensitivity of a PHQ-9 score greater than 10 is 88 percent
        - Specificity of a PHQ-9 score greater than 10 is also 88 percent
        - ROC analysis: area under the curve for the PHQ-9 in diagnosing major depression was 0.95
    - The PHQ-9M was not independently tested for validity.
- Validity Testing Conducted at the Patient/Encounter Level:
    - The developer also presents empirical encounter-level validity testing by analyzing the results of their standard data quality checks and audits. These checks are done on (1) date of birth, (2) date of service, (3) icd-10 codes used, (4) attestation of inclusion of patients, (5) exclusions to the measure.
        - 49% of groups passed with no errors; 58% of those that submitted data passed initial quality checks; 30% of groups that submitted data were audited; 94% passed the audit.
        - Results on testing of all critical data elements is not provided; therefore, this does not meet NQF criteria for sufficient critical data element testing.
- Validity Testing Conducted at the Accountable Entity Level:
    - Correlation was performed against a depression outcome measure to test the hypothesis that clinics that do well assessing their patients with a diagnosis of depression frequently with the PHQ-9/PHQ-9M will also perform better in achieving remission (PHQ-9<5) at six months.
        - Adults: the correlation between assessment with PHQ-9/PHQ-9M and Depression Remission at Six months was R-squared = 0.1754.
        - Adolescents: the correlation between assessment with PHQ-9/PHQ-9M and Depression Remission at Six months was R-squared = 0.2744
        - Both stratifications demonstrate fairly weak positive correlations against a theoretically-related outcome measure. However, the developer states that it is important to continually assess patients with a current diagnosis of depression or a history of depression for depression symptoms as the measure supports the outcome measures of depression remission and response at six and twelve months. (NQF # 0710, 0711, 1884 and 1885)

Exclusions

- Exclusions are of a clinical nature and include those for whom outcomes may be different due to life expectancy (hospice, nursing home resident, death) or co-morbid diagnoses that emerge after initial diagnosis of depressive disorder.
- Exclusions occurred at a rate of 3.45 percent, the highest categories of which were other co-morbid disorders (schizophrenia, bipolar, personality, pervasive developmental).
- Developer states that overall, exclusions do not limit or reduce the desired target population of patients with major depression or dysthymia, and that the updated analysis demonstrates continued appropriate clinical indication without reducing the target population, as concurrent diagnose can have very different outcomes.

Risk-Adjustment

- The measure is not risk adjusted or stratified.

Meaningful Differences

- The developer states that MN statewide averages are 77.7 percent for adults and 78.4 percent for adolescent populations. The range of results for adults was 25 to 100 percent, and for adolescents was eight to 100 percent.
    - The developer provides box plots to show the wide variability in medical group rates and states that for adults, a large portion of clinics are in the lower quartile.
    - The developer states that these data indicate that this measure can identify meaningful differences and continues to demonstrate opportunity for improvement.

Missing Data

- The developer states that missing data are not an issue for this measure as patients with a diagnosis of major depression or dysthymia who have a visit or contact within the measurement period who are not assessed at least once in the four-month period remain in the denominator.
- This measure is a companion related measure that allows medical groups to understand their use of the PHQ-9/ PHQ-9M tool in assessing depression and related to remission and response outcome measures (NQF #s 0710, 0711, 1884 and 1885)

Comparability

- The measure only uses one set of specifications.

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*

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

Committee Pre-evaluation Comments:

2a. Reliability

- 2a1. Reliability-Specifications

    - specifications are clear. Recent updates to this and other related measures makes it difficult to assess whether there is appreciable improvement over time, which is a requirement for maintenance of a measure?
    - Concerns about the way PHQ9 completion is recorded. Not all EMR systems may record uniformly or adequately; therefore it may be difficult to accurately capture completion of this metric.
    - No concerns regarding reliability specifications.
    - No concerns, clearly defined.
    - Reliable
    - Data elements clearly defined and descriptors provided. All steps are clear. No concerns about measure being implemented consistently.
    - Data elements are well defined and clear and measure has been consistently implemented.
- 2a2. Reliability-Testing
    - Given aggregated data, the team is left with beta binomial model to statistically explore capacity to distinguish higher vs. lower performing clinics. Studies supporting the psychometric properties of the PHQ 9 support the selection of this screener but not really the reliability and validity of the quality measure.
    - PHQ-9M should undergo reliability testing.
    - No
    - No concerns.
    - No
    - No
    - No

2b. Validity

- validity testing based on their own internal quality checks and audits which is really assessing data quality internally. A bit of a leap to explore correlation to adherence to dep remission measure with weak positive correlations, but part of the problem may also be that for the remission measures persons with no screener during follow-up time window were counted as non-remitters--so just not strong validity testing.
- N/A
- Some issues with validity testing - including at patient/encounter level (results on testing at all levels not provided) and at accountable entity level (both stratifications demonstrate fairly weak positive correlations against a theoretically-related outcome measure).
- No concerns.
- No
- No. Discuss the missing testing on critical data elements.
- No

2b2-2b6. Potential threats to validity

- 2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)
    - not risk adjusted because the data source was reported EHR data from clinics or medical groups, just yes/no within the calendar quarter. No patient level of data on clinical severity, no capacity to identify types of treatment or where in episode of care for the person.

- Measure should be risk stratified because there are well-known disparities in depression care for certain racial/ethnic groups. Also access to care could greatly affect the administration of the PHQ9
- Clinical exclusions are well justified; no risk adjustment.
- The measures exclusions consisted of comorbid disorders- appropriate. The measure is not risk adjusted.
- Acceptable
- Exclusions are appropriate. Measure not risk adjusted or stratified. Developer includes information comparing medical group rates and states that there are meaningful differences, speaks to the measures ability to identify opportunities for improvement. Only one set of specifications. The developer explains why missing data will not impact validity.
- This seems adequate

- 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data)

Criterion 3. Feasibility

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

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

- PHQ-9 may be regularly captured in the course of care for clinical purposes which is the subject of this quality measure. However, it may not be a standard part of care in many settings, which this measure is trying to change.

  MNCM developed a direct data submission process in 2006 whereby medical groups submit a patient level data file of a minimal data set (only those elements needed for measure calculation, risk adjustment and stratification/ analysis) to their HIPAA secure data portal for rate calculation and public reporting. The developer has provided additional findings from this process.
- The developer notes that MNCM is implementing a new data collection method, PIPE (Process Intelligence Performance Engine) that serves as a warehouse of clinical data (encounters, problem lists, labs, medications, etc) where measures are calculated centrally, significantly reducing data collection burden for providers.
- This measure was originally developed as an e-CQM (legacy measure) and one of the first adopted into CMS' Measure Authoring Tool (MAT), CMS 160 8.4 and was used for several years in the e-CQM program until it was recommended for removal from the MIPS program by CMS as part of the 2020 rule making process (effective MIPS Payment Year 2022). Rationale indicated favor for the more robust companion outcome measure Depression Remission at Twelve Months (Q370/CMS159/NQF 0710e).
- There are no fees associated with this measure. The PHQ-9 is publicly available for use.

Questions for the Committee:

- *Is the data collection strategy ready to be put into operational use?*

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

Committee Pre-evaluation Comments:

3. Feasibility

- The developer appropriately states that the PHQ 9 "may not be a standard part of care in many settings, which this measure is trying to change." The conundrum is they are advocating for use of one depression screener. This is inconsistent with The Joint Commission that allows a pool of validated measures for suicide screening, and the Core Set measures re: dep screening and follow-up (developer CMS) that allows for a pool of screening measures.
- The ability to capture PHQ9 administration.
- No concerns about feasibility.
- concern - is the data readily available
- Feasible
- Performance results are reported back annually to groups who submit data. Publicly reported on the MN Health Scores.
- feasibility is good

Criterion 4: Use and Usability

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 evaluates 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
Planned use in an accountability program?	☐	Yes	☐	No	☒	NA

Accountability program details

- The measure is reported on MN Community Measurement- a non-profit 501 (c)(3) whose mission is to accelerate the improvement of health by publicly reporting health care information.
- The measure is used in all primary care clinics in MN and bordering communities in Wisconsin, North Dakota, South Dakota and Iowa.
- The measure is reported publicly on MN HealthScores, a consumer facing public reporting website.
- Rates for this measure are published annually in the MNCM Health Care Quality Report.

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

feedback on the measure performance or implementation; 3) this feedback has been considered when changes are incorporated into the measure

Feedback on the measure by those being measured or others

- The developer provides results to measured entities and allows measure users to appeal results prior to public reporting.
- Measure Review Committee and many medical groups identify challenges with the technical replication of this measure in the medical group's internal systems (index event and follow-up window and the difficulty in maintaining ongoing contact with patients who are depressed).
- Periodically, MNCM surveys all medical groups in MN to assess value in measures and feasibility/ease/difficulty in data collection and submission to MNCM for measure rate calculation. Ease/difficulty ratings for the depression measures improved by 9 percent. 57 percent of those surveyed rated the depression measures as high/moderate value.
- Feedback was used in the decision to add adolescents to this measure.

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

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

Improvement results

- The developer provides data that show gradual improvement in this measure from 55.4 percent to 77.7 percent from 2010 to 2020. This data incorporates years prior to re-specification of the measure in 2020 (dates of index event 1/1/2018 to 12/31/2018) and of the measure as specified in the submission.
- The developer also notes that the denominator of eligible patients has grown from 108,261 in 2010 to over 244,00 in 2020 and that this demonstrates increased screening for depression.

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

- The developer does not identify any unintended negative consequences.
- The developer notes that incorporating adolescents into the measure may help address MDD early and aid in prevention over the life cycle.

Potential harms

- No potential harms identified by the developer.

***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:

4a. Use

- adherence rates are publicly reported and stakeholder input provided
- Yes. Would be important to determine how this can occur at a national level.
- Public reporting and use in accountability programs, along with active solicitation of feedback, indicative of appropriate accountability and transparency.
- Yes feedback opportunities were provided and implemented to improve the measure. The measure is useful for decision making.
- May affect performance in systems prepared to use the information accordingly
- Groups being measured have been given their results and assistance in interpreting the data. Feedback has been gathered and is considered when implementing changes.
- It's use is growing

4a. Usability

- a concern at the provider level is that the screener could place them at risk for medical liability if detect SI but don't have the resources to connect patient with urgent psychiatric care/evaluation
- Benefits outweigh the risks.
- Gradual improvement from 2010 to 2020 suggest continued improvement. Benefits appear to outweigh harms, without any apparent unintended consequences.
- The Benefits outweigh any potential harm. By including adolescents early intervention can lead to treatment that will enhance quality of life. There are no harms.
- Not clear that we have data that has systematically looked at this issue (e.g., undertreated depression leading to more treatment resistance)
- Tracking scores over time can assist in the development and use of effective interventions for treatment of depression. It is noted that there have been gradual improvements in the measure over time. The feedback to specific sites allows them the opportunity for quality improvement of current treatments. No unintended consequences and ongoing measurement should have significant benefits to patient populations in the treatment of depression, especially with more opportunities for early detection with including adolescents.
- we can see that improvement is occurring and there are no direct harms

Criterion 5: Related and Competing Measures

Related measures

- 1885: Depression Response at Twelve Months- Progress Towards Remission
- 0710e: Depression Remission at Twelve Months

- 0711: Depression Remission at Six Months
- 1884: Depression Response at Six Months- Progress Towards Remission

Harmonization:

- The developer states that these related measures are all harmonized with this measure.

Committee Pre-evaluation Comments:

5: Related and Competing Measures

- see prior comments re: CMS Dep/Screening measures in Core Set and Joint Commission accreditation requirements
- There are related measures being used by NCQA.
- Four related measures; a little surprised not to see more elaboration re: harmonization.
- No competing measures. Related measures- • 1885: Depression Response at Twelve Months- Progress Towards Remission • 0710e: Depression Remission at Twelve Months • 0711: Depression Remission at Six Months • 1884: Depression Response at Six Months- Progress Towards Remission
- YES A BOATLOAD This is my major concern.
- All related measures have been harmonized.
- No competing measures. It is a necessary component of the following related measures:1885: Depression Response at Twelve Months- Progress Towards Remission • 0710e: Depression Remission at Twelve Months • 0711: Depression Remission at Six Months • 1884: Depression Response at Six Months- Progress Towards Remission

Public and NQF Member Comments (Submitted as of June 16, 2022)

Member Expression of Support

- Of the 1 NQF members who have submitted a expression of support, 0 expressed “support” and 1 expressed “do not support” for the measure.

Comments

Comment 1 by: Submitted by Collette Cole, Minnesota Community Measurement

Hello, MN Community Measurement is submitting this comment in response to NQF staff feedback about insufficient evidence for this measure #0712 Depression Assessment with the PHQ-9/PHQ9-M. It was noted that there was no empirical evidence to demonstrate that performing the PHQ-9 in and of itself in isolation is associated with any improvement in outcomes. In terms of measurement and assessing outcomes, frequent and ongoing assessment with the PHQ-9/PHQ-9M is key to understanding the patient’s progress towards the reduction of depression symptoms. Administering the PHQ-9 is like taking a blood pressure- you need to do something with the information to affect the outcome of hypertension. Depression is now being considered the sixth vital sign by many and assessing patients is critical to identifying depression and improving outcomes. [Trivedi, M., Jha, M. et al VitalSign6: a Primary Care First (PCP-First) Model for Universal Screening and Measurement-Based Care for Depression. Pharmaceuticals 2019, 12, 71; doi:10.3390/ph12020071] Supportive evidence was provided in the context of the Institute for Clinical Systems Improvement Depression Care guidelines for 1) Comprehensive Treatment Plan with Shared Decision Making- Collaborative Care Model and 2) Establish Follow-up Plan. In addition, PubMed lists 14,699 studies associated with the use of the PHQ-9 for measuring or monitoring depression including a 2021 meta-analysis that supports the use of this tool to determine outcomes [Negeri, Z.F., Levis, B., Sun, Y. et al Accuracy of the Patient Health Questionnaire-9 for screening to detect major

depression: updated systemic review and individual participant data meta-analysis BMJ 2021 Oct 5;375:n2183. Doi:10.1136/bmj.n.2183]. This measure is an important companion to outcome measures of response and remission, serving two purposes: the first is to understand how well a practice does at assessing their patients who have a diagnosis of depression, and the second is to guard against gaming of the outcome measures through selective administration of the PHQ-9. Sincerely, Collette Cole, RN BSN CPHQ Clinical Measure Developer, MN Community Measurement

Comment 2 by: Submitted by Koryn Rubin, American Medical Association

The American Medical Association (AMA) appreciates the opportunity to comment on this measure. We are writing to request clarification on one item with this measure. We seek clarification on whether this measure is intended to be captured as an electronic clinical quality measures (eCQMs) since the complimentary measure (710e Depression Remission at Twelve Months), which is an eCQM, uses the same data and is specified similarly. It would seem counterintuitive to have related measures endorsed that leverage what appear to be the same data, yet are endorsed with different data sources and specifications. If it is intended to be an eCQM, our concerns on the inadequate testing and missing feasibility scorecard for NQF #710e would also apply to this measure. The AMA requests clarification on whether the measure is intended to be an eCQM be addressed prior to continued endorsement of this measure. We appreciate the Committee's consideration of our comments.

Comment 3 by: Submitted by Steven Inman

Dear NQF: I represent Children's Health Network, the Network affiliated with Children's Hospitals and Clinics of Minnesota, and I support the re-endorsement of the suite of depression measures currently under review by the Behavioral Health Standing Committee. We have been using these measures for many years at our organization to understand and support positive care and outcomes for patients with depression. Additionally we have pay-for-performance contracts with insurance payers and recognition programs that utilize the rates for these measures. Using the PHQ-9 helps clinics in screening, diagnosing and ongoing monitoring of symptoms of depression. Our organization has increased focus on depression care and value these measures that support our focus. The outcome measures, work together in measuring outcomes at multiple points in time using the same information to measure remission or progress towards remission. The use of this measures on a statewide basis in Minnesota helps to focus attention on these outcomes for an important health problem that impacts many people. I support the continued endorsement of all five measures (NQF#s 0710e, 0711, 0712, 1884 and 1885). Respectfully; Steven Inman, MD Pediatrician Medical Director - Children's Health Network of Minnesota

NQF Staff Scientific Acceptability Evaluation

RELIABILITY: SPECIFICATIONS

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

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

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

2. **Briefly summarize any concerns about the measure specifications.**
   - There are no concerns with the measure specifications
   - Specifications have been updated since last endorsement. Changes include:
     - Incorporating adolescents ages 12-17 into the measure

- Adding PHQ-9M PRO tool to the measure (modified for teens)
- Modification of the exclusion value set of personality disorder
- Addition of exclusions for schizophrenia and pervasive developmental disorder

RELIABILITY: TESTING

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

Testing Population

- This measure is in full implementation and thus includes data from all primary care and behavioral health (psychiatry) clinics in Minnesota. Due to the condition's chronic episodic nature, no sampling is allowed and the full population of eligible patients, regardless of payer, is included.
  - Sites represent all primary care and behavioral health (psychiatry) clinics in Minnesota and bordering cities in other states that wish to participate. Clinics represent urban and rural, large multi-specialty health care systems, medium and small practices that care for adult patients with depression. 103 medical groups representing 615 clinics were included in the testing of this measure, representing 227,127 adults and 12,616 adolescents.
  - Testing used adult patients with dates of service 10/1/2019 to 1/31/2020 reported in 2020, and adolescent patients age 12 to 17 with dates of service 1/1/2019 to 12/31/2020.
  - Clinics must have greater than or equal to 30 patients in the denominator to be included.

Reliability of the PHQ-9

- Reliability of the PROM has been validated in the literature in the adult population, including calculating a Cronbach's alpha and test-retest reliability.
  - The PHQ-9 has also been validated in adolescent populations (ages 13 to 17), but the PHQ-9M Modified for Teens has not undergone separate validation studies. The developer attests that this version has essentially the same 9 questions as the PHQ-9 with slight wording variation for an adolescent population.

Reliability of the Measure Score

- Empirical testing of computed performance scores for reportable clinics was conducted using a beta-binomial model (signal-to-noise).

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

   Submission document: Testing attachment, section 2a2.3

Reliability of the PROM

- Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study
- Cronbach's alpha of 0.86 in the PHQ OBGYN Study
- Test-retest showed the correlation between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours was 0.84, and the mean scores were nearly identical (5.08 vs 5.03).
- The developer states that this testing demonstrates the PHQ-9 tool is appropriate for measuring patient outcomes related to depression.

Reliability of the measure score:

- The developer found that the using 601 clinics and 227,000 patients, the PHQ-9 Assessment- Adults had an average reliability score of 0.932903. IQR was not provided.
- The developer found that using 142 clinics and 12,616 patients, the PHQ-9 Assessment- Adolescents had an average reliability score of 0.878959. IQR was not provided.
- The developer states that a beta-binomial reliability score of greater than 0.70 indicates that it is acceptable to draw conclusions about groups, in this case there is the ability to distinguish higher performing clinics from lower performing clinics.

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.**

    Specifications are precise (Box 1) -> Testing conducted with the measure as specified (Box 2) -> Testing conducted at the accountable entity level for specified level of analysis (Box 4) -> Method was appropriate (Box 5) -> There is moderate certainty scores are reliable (PHQ-9M was not separately tested) (Box 6b) -> Rate as MODERATE

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

Submission document: Testing attachment, section 2b2.

- Exclusions are of a clinical nature and include those for whom outcomes may be different due to life expectancy (hospice, nursing home resident, death) or co-morbid diagnoses that emerge after initial diagnosis of depressive disorder.
- Exclusions occurred at a rate of 3.45%, the highest categories of which were other co-morbid disorders (schizophrenia, bipolar, personality, pervasive developmental).
- Developer states that overall, exclusions do not limit or reduce the desired target population of patients with major depression or dysthymia, and that the updated analysis demonstrates continued appropriate clinical indication without reducing the target population, as concurrent diagnose can have very different outcomes.
- No concerns.

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

Submission document: Testing attachment, section 2b4.

- adolescent populations. The range of results for adults was 25 to 100 percent, and for adolescents was eight to 100 percent.
    - The developer provides box plots to show the wide variability in medical group rates and states that for adults, a large portion of clinics are in the lower quartile.
- The developer states that these data indicate that this measure can identify meaningful differences and continues to demonstrate opportunity for improvement.

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.

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

Submission document: Testing attachment, section 2b6.

- The developer states that missing data is not an issue for this measure as patients with a diagnosis of major depression or dysthymia who have a visit or contact within the measurement period who are not assessed at least once in the four-month period remain in the denominator.
- This measure is a companion related measure that allows medical groups to understand their use of the PHQ-9/ PHQ-9M tool in assessing depression and related to remission and response outcome measures (NQF #s 0710, 0711, 1884 and 1885)

16. **Risk Adjustment**

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

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

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

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

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

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

16d.**Risk adjustment summary:**

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

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

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

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

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

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

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

☐ Yes ☐ Somewhat ☐ No (If "Somewhat" or "No", please explain)

18. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**

VALIDITY: TESTING

19. **Validity testing level: ☐ Measure score ☐ Data element ☒ Both**

20. **Method of establishing validity of the measure score:**

☐ Face validity

☒ Empirical validity testing of the measure score

☐ N/A (score-level testing not conducted)

21. **Assess the method(s) for establishing validity**

Submission document: Testing attachment, section 2b2.2

- Validity Testing Conducted at the Patient/Encounter Level:
  - The developer references testing of construct validity in the literature, using mental health professional re-interview as the criterion standard
  - The developer also presents empirical encounter-level validity testing by analyzing the results of their standard data quality checks and audits. These checks are done on (1) date of birth, (2) date of service, (3) icd-10 codes used, (4) attestation of inclusion of patients, (5) exclusions to the measure.

- Validity Testing Conducted at the Accountable Entity Level:
  - Correlation was performed against a depression outcome measure to test the hypothesis that clinics that do well assessing their patients with a diagnosis of depression frequently with the PHQ-9/PHQ-9M will also perform better in achieving remission (PHQ-9<5) at six months.

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

Submission document: Testing attachment, section 2b2.3

Testing of the PROM PHQ-9

- In addition to the adults and elderly, the PHQ-9 has been validated in the adolescent populations (age 13 to 17).
    - The developer states that PHQ-9M is only a slight modification of the original tool as the nine questions are essentially the same as the original PHQ-9, which has been validated for adolescents ages 13 and older. The APA recommends using the modified version of the PHQ-9 for children ages 11 to 17 to assess depression symptom severity (APA, 2015) and does not have separate validity testing results.

Empirical Testing at the Accountable Entity Level

- Adults: the correlation between assessment with PHQ-9/PHQ-9M and Depression Remission at Six months was R-squared = 0.1754.
- Adolescents: the correlation between assessment with PHQ-9/PHQ-9M and Depression Remission at Six months was R-squared = 0.2744
    - Both stratifications demonstrate fairly weak positive correlations against a theoretically-related outcome measure. However, the developer states that it is important to continually assess patients with a current diagnosis of depression or a history of depression for depression symptoms as the measure supports the outcome measures of depression remission and response at six and twelve months. (NQF # 0710, 0711, 1884 and 1885).

23. **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)

24. **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)

25. **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.)

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

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **What is the level of certainty or confidence that the empirical analysis demonstrates that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct?**

    ☐ **High**

    ☐ **Moderate**

    ☐ **Low**

    ☐ **Insufficient**

28. **Briefly explain rationale for rating of EMPIRICAL ANALYSES TO SUPPORT COMPOSITE CONSTRUCTION**

ADDITIONAL RECOMMENDATIONS

29. **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.**

Criteria 1: 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

1ma.01. Indicate whether there is new evidence about the measure since the most recent maintenance evaluation. If yes, please briefly summarize the new evidence, and ensure you have updated entries in the Evidence section as needed.

[Response Begins]
Yes
[Yes Please Explain]
Since the last maintenance endorsement of this measure, the age range was expanded to include adolescents. Evidence related to adolescents was added to evidence submitted previously.

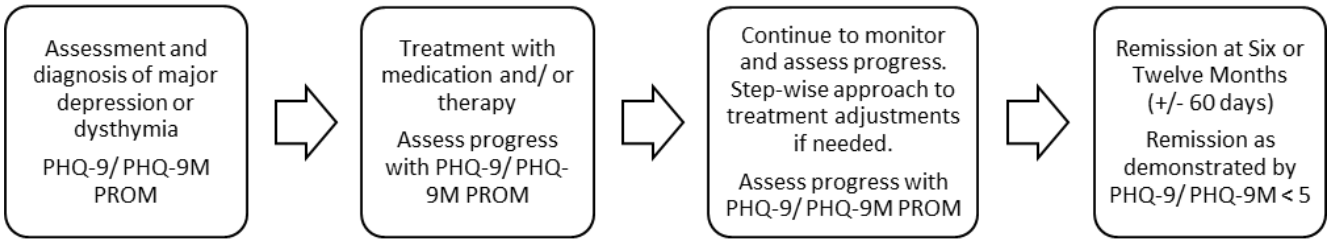
[Response Ends]

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:
2021 Submission:
Updated evidence information here.
2018 Submission:
Evidence from the previous submission here.

1a.01. Provide a logic model.

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.

[Response Begins]
2021 Submission
Updates reflect the addition of the PHQ-9M tool for adolescents



Health Care Process Steps to Achieve Desired Outcome for Depression

Please note that this process measure for administration of the PHQ-9/PHQ-9M depression tool, a PROM that is validated for both the assessment and diagnosis of depression as well as for monitoring ongoing outcomes of treatment, is a related process measure that supports outcome measures of depression remission (PHQ-9/PHQ-9M < 5) and depression response (PHQ-9/PHQ-9M is improved by ≥ 50%) at six and twelve months. To quote a NQF Behavioral Steering Committee member, as these measures were initially endorsed, "the best way to avoid being measured is to never give the PHQ-9". This process measure allows an understanding of the use of the tool in the target population, promotes frequent and follow-up contact with patients whose score indicates a need for treatment and serves as a catalyst in a

collaborative care model for patients with major depression or dysthymia. It is estimated that up to 90% of patients diagnosed with depression and anxiety are treated solely in primary care. [NICE National Institute Health and Care Excellence United Kingdom 2011]

Severity	PHQ-9 Scores	Possible Diagnoses	Treatment Recommendations
Undefined	Initial Score: 5-9	Does not meet criteria for major depressive disorder	Consider for persistent depressive disorder Stay in touch: a) If no improvement after one or more months, consider treating or referral to behavioral health. b) If symptoms deteriorate, start treatment or make a referral.
*	Follow-up Score: 5-9	Partial remission	Continue stepped therapies approach.
Per DSM-5: Few, if any, symptoms in excess of those required to make the diagnosis are present, the intensity of the symptoms is distressing but manageable, and the symptoms result in minor impairment in social or occupational functioning.	10-14	Mild major depression	Combined psychotherapy and pharmacotherapy treatment. When unable to do combined therapy due to patient preferences, availability and affordability of the treatments, start with psychotherapy. Initially consider weekly contacts to ensure adequate engagement, then at least monthly.
Per DSM-5: The number of symptoms, intensity of symptoms, and/or functional impairment are between those specified for "mild" and "severe."	15-19	Moderate major depression	Combined psychotherapy and pharmacotherapy treatment. When unable to do combined therapy due to patient preferences, availability and affordability of the treatments, start with psychotherapy. Initially consider weekly contacts to ensure adequate engagement, then at least every 2-4 weeks.
Per DSM-5: The number of symptoms is substantially in excess of that required to make the diagnosis, the intensity of the symptoms is seriously distressing and unmanageable, and the symptoms markedly interfere with social and occupational functioning.	≥20	Severe major depression	Combined psychotherapy and pharmacotherapy treatment. When unable to do combined therapy due to patient preferences, availability and affordability of the treatments, start with pharmacotherapy. Weekly contacts until less severe.
Meets DSM-5 criteria for persistent depressive disorder	*	Pure dysthymia	Consider starting with medication. Consider stepped care, which includes augmenting medications and adding psychotherapy for patients who don't improve.
Meets DSM-5 criteria for persistent depressive disorder	*	Chronic major depression	Combined psychotherapy and pharmacotherapy treatment.

Institute for Clinical Systems Improvement Clinical Practice Treatment Guidelines

* Cell intentionally left empty

[Response Ends]

1a.02. Select the type of source for 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.

[Response Begins]

Clinical Practice Guideline recommendation (with evidence review)
[Response Ends]

If the evidence is not based on a systematic review, skip to the end of the section and do not complete the repeatable question group below. If you wish to include more than one systematic review, add additional tables by clicking "Add" after the final question in the group.

Evidence - Systematic Reviews Table (Repeatable)
Group 1 - Evidence - Systematic Reviews Table

1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]
Guidelines for Adults
Institute Clinical Systems Improvement Depression in Primary Care Guideline
Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, Mack N, Myszkowski M. Institute for Clinical Systems Improvement. Adult Depression in Primary Care. Updated March 2016.
https://www.icsi.org/wp-content/uploads/2019/01/Depr.pdf

[Response Ends]

1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]
The PHQ-9 has been validated for measuring depression severity (Kroenke, 2001; Spitzer, 1999) and is validated as a tool for both detecting and monitoring depression in primary care settings (Kroenke, 2010; Wittkampf, 2007). It has a sensitivity (false negative) of 0.77 and specificity (false positive) of 0.85 when using the screened item scoring method. Two other tools with good utility in case finding, aiding diagnosis and severity grading are the Structural Clinical Interview DSM-IV Axis-I Disorders (SCID-I) with a sensitivity of 85% and specificity 82% and the Mini International Neuropsychiatric Interview (MINI) with a sensitivity of 78% and specificity of 85% (Pettersson, 2015). [page 16]
Discuss Treatment Recommendations Primary goal. When considering treatment options, the primary goal is to achieve remission or to get the patient to be predominately symptom-free (i.e., a PHQ-9 score of less than five) (Kroenke, 2001). [page 30]
PHQ-9 as monitor and management tool. The PHQ-9 is an effective management tool, as well, and should be used routinely for subsequent visits to monitor treatment outcomes and severity. It can also help the clinician decide if/how to modify the treatment plan (Duffy, 2008; Löwe, 2004). Using a measurement-based approach to depression care, PHQ-9 results and side effect evaluation should be combined with treatment algorithms to drive patients toward remission. A five-point drop in PHQ-9 score is considered the minimal clinically significant difference (Trivedi, 2009). [page 50]

[Response Ends]

1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]
The Institute for Clinical Systems Improvement uses a GRADE methodology to rate evidence and strength of recommendation. The definitions for evidence and relationship to the strength of recommendation are located in the full methodology explanation in 1a.06. There are two recommendations within this guideline that support the ongoing assessment of symptoms for patients with depression.
6. Comprehensive Treatment Plan with Shared Decision-Making Collaborative Care Model
Recommendation: A collaborative care approach is recommended for patients with depression in primary care.
Quality of Evidence and Strength of Recommendation: Quality of Evidence: High Strength of Recommendation: Strong
Benefit: Collaborative care model has demonstrated improvement in treatment adherence, patient quality of life and depression outcomes. It has demonstrated beneficial impact on direct and indirect economic benefits. Evidence suggests the collaborative care model is also effective for depression during pregnancy and postpartum period.

Harm: There are challenges in providing the collaborative care model, such as identifying depressed patients, identifying care managers with the right experience and background, establishing the responsibilities and scope of practice of the care managers, whether to locate care managers in a clinic vs. centrally based, determining the level of psychiatric supervision, seeking adequate reimbursement for services provider to ensure program sustainability, and feasibility of small clinics to employ on-site mental health specialists or fulltime care managers.
Benefit-Harms Assessment: Collaborative care has shown to improve patient outcomes and provider satisfaction while decreasing cost outweighing the challenges of implementing a collaborative care program.
Relevant Resources: Fortney, 2013; Archer, 2012; Katon, 2008; Gjerdingen, 2007; Belnap, 2006; Gilbody, 2006; Hunkeler, 2006; Simon, 2001a; Katon, 1999
Randomized controlled trials have demonstrated the effectiveness of the collaborative care model, in which primary care treatment of depression is provided by a team (depression care manager, primary physician, consulting psychiatrist and others). The work group recommends three key references (Gilbody, 2006; Hunkeler, 2006; Katon, 1999). This model has demonstrated improvement in treatment adherence, patient quality of life and depression outcomes (Archer, 2012; Gilbody, 2006; Hunkeler, 2006; Katon, 1999).
Beneficial impact on direct medical costs can also be found. Further dissemination of this model has been recommended (Simon, 2001a). Katon, 2008 summarizes and solidifies the argument for collaborative care in the treatment of depression, the direct and indirect economic benefits of collaborative care, as well as improved outcomes (Katon, 2008). Evidence suggests the collaborative care model is also effective for depression during pregnancy and postpartum (Gjerdingen, 2007).
Improved Patient Outcomes
Better medication compliance and reduced risk of relapse. The use of a collaborative care model can help with medication compliance by providing closer follow-up than is possible without a care manager. Three or more follow-up visits in the first three months reduced the risk of relapse/recurrence of depression, as did continuous use of antidepressants (Kim, 2011). Care management facilitates continuous use of antidepressants by providing close follow-up and early intervention when side effects occur.
Reduced suicidal ideation
In the Prevention of Suicide in Primary Care Elderly: Collaborative Trial (PROSPECT) study, suicidal ideation rates declined in patients receiving care based on treatment guidelines and use of a care manager (Bruce, 2004). In the Improving Mood Providing Access to Collaborative Treatment (IMPACT) study, 1,801 primary care patients were randomly assigned to collaborative care or usual care. Intervention subjects had less suicidal ideation at 6 and 12 months, and there were no completed suicides for either group in 18 months (Unützer, 2006).
Improved Provider Satisfaction
The rewards for health care organizations that implement collaborative care models for their depressed patients are substantial, not only for the patients, but also for physician satisfaction. Of physicians participating in the IMPACT trial (Levine, 2005), only 54% were satisfied with the resources they had to treat depressed patients before the trial. This satisfaction was independent of practice setting (fee-for-service versus capitated). Sixty-four percent of physicians self-rated their ability to provide at least "very good" depression care before IMPACT. Eighty-five percent of clinicians before IMPACT felt that a collaborative care model would be helpful in treating patients with depression, diabetes or heart failure (Levine, 2005). Afterwards, 90% of physicians described the collaborative care program as helpful in treating patients with depression. Ninety-three percent of physicians were at least somewhat satisfied with the resources available for treating depressed patients assigned to the IMPACT model, whereas only 61% were somewhat satisfied if their patients were assigned to usual care (Levine, 2005). Ninety-four percent of clinicians rated the care managers as somewhat or very helpful in treating depression, and 82% indicated that IMPACT program improved their patients' clinical outcomes. Clinicians identified the two most helpful features of the program as "proactive patient follow-up" and "patient education" (Levine, 2005).
High Quality Evidence: Further research is very unlikely to change our confidence in the estimate of effect.
Strong Recommendation: The work group is confident that the desirable effects of adhering to this recommendation outweigh the undesirable effects. This is a strong recommendation for or against. This applies to most patients.
7a. Establish Follow-Up Plan
Recommendation: Clinicians should establish and maintain follow-up with patients.
Quality of Evidence and Strength of Recommendation: Quality of Evidence: Low Strength of Recommendation: Strong
Benefit: Appropriate, reliable follow-up is highly correlated with improved response and remission scores. It is also correlated with the improved safety and efficacy of medications and helps prevent relapse.
Harm: Potential harms may include added expense and unnecessary visits.
Benefit-Harms Assessment: Benefits appear to outweigh potential harms by a wide margin
Relevant Resources: Trivedi, 2006b; Unützer, 2002; Hunkeler, 2000; Simon, 2000

Proactive follow-up contacts (in person, telephone) based on the collaborative care model have been shown to significantly lower depression severity (Unützer, 2002). In the available clinical effectiveness trials conducted in real clinical practice settings, even the addition of a care manager leads to modest remission rates (Trivedi, 2006b; Unützer, 2002). Interventions are critical to educating the patient regarding the importance of preventing relapse, safety and efficacy of medications, and management of potential side effects. Establish and maintain initial follow-up contact intervals (office, phone, other) (Hunkeler, 2000; Simon, 2000).
PHQ-9 as monitor and management tool. The PHQ-9 is an effective management tool, as well, and should be used routinely for subsequent visits to monitor treatment outcomes and severity. It can also help the clinician decide if/how to modify the treatment plan (Duffy, 2008; Löwe, 2004). Using a measurement-based approach to depression care, PHQ-9 results and side effect evaluation should be combined with treatment algorithms to drive patients toward remission. A five-point drop in PHQ-9 score is considered the minimal clinically significant difference (Trivedi, 2009). Every time that the PHQ-9 is assessed, suicidality is assessed, as well. If the suicidality was indeed of high risk, urgent referral to crisis specialty health care is advised. In case of low suicide risk, the patient can proceed with treatment in the primary care practice (Huijbregts, 2013).
Collaboration with Mental Health Consider collaborating with a behavioral health care clinician for the following: • Patient request for psychotherapy • Presence of severe symptoms and impairment in patient, or high suicide risk • Presence of other psychiatric condition (e.g., personality disorder or history of mania) • Suspicion or history of substance abuse • Clinician discomfort with the case • Medication advice (psychiatrist or other mental health prescriber) • Patient request for more specialized treatment
Low Quality Evidence: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain.
Strong Recommendation: The work group feels that the evidence consistently indicates the benefit of this action outweighs the harms. This recommendation might change when higher quality evidence becomes available.
https://www.icsi.org/wp-content/uploads/2021/11/Depr.pdf

[Response Ends]

1a.06. Provide all other grades and definitions from the evidence grading system.

[Response Begins]
GRADE Methodology ICSI utilizes the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology system. GRADE has advantages over other systems including the former system used by ICSI. Advantages include:

- development by a widely representative group of international guideline developers;
- explicit and comprehensive criteria for downgrading and upgrading quality of evidence ratings;
- clear separation between quality of evidence and strength of recommendations that includes a transparent process of moving from evidence evaluation to recommendations;
- clear, pragmatic interpretations of strong versus weak recommendations for clinicians, patients and policy-makers; • explicit acknowledgement of values and preferences; and
- explicit evaluation of the importance of outcomes of alternative management strategies. GRADE involves systematically evaluating the quality of evidence (high, moderate, low, very low) and developing a strength of recommendation (strong, weak). For more detailed information on GRADE, please go to: http://www.gradeworkinggroup.org/. GRADE Methodology definitions:

Category	Quality Definitions	Strong Recommendation	Weak Recommendation
High Quality Evidence	Further research is very unlikely to change our confidence in the estimate of effect.	The work group is confident that the desirable effects of adhering to this recommendation outweigh the undesirable effects. This is a strong recommendation for or against. This applies to most patients.	The work group recognizes that the evidence, though of high quality, shows a balance between estimates of harms and benefits. The best action will depend on local circumstances, patient values or preferences.

Category	Quality Definitions	Strong Recommendation	Weak Recommendation
Moderate Quality Evidence	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.	The work group is confident that the benefits outweigh the risks but recognizes that the evidence has limitations. Further evidence may impact this recommendation. This is a recommendation that likely applies to most patients.	The work group recognizes that there is a balance between harms and benefits, based on moderate quality evidence, or that there is uncertainty about the estimates of the harms and benefits of the proposed intervention that may be affected by new evidence. Alternative approaches will likely be better for some patients under some circumstances.
Low Quality Evidence	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain.	The work group feels that the evidence consistently indicates the benefit of this action outweighs the harms. This recommendation might change when higher quality evidence becomes available.	The work group recognizes that there is significant uncertainty about the best estimates of benefits and harms.

https://www.icsi.org/wp-content/uploads/2021/11/Depr.pdf

[Response Ends]

1a.07. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]
Recommendation 6: Collaborative care approach is recommended for patients with depression in primary care.
High Quality Evidence: Further research is very unlikely to change our confidence in the estimate of effect.
Strong Recommendation: The work group is confident that the desirable effects of adhering to this recommendation outweigh the undesirable effects. This is a strong recommendation for or against. This applies to most patients.
Recommendation 7a: Establish follow-up plan. Use of PHQ-9 as monitor and management tool
Low Quality Evidence: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change. The estimate or any estimate of effect is very uncertain.
Strong Recommendation: The work group feels that the evidence consistently indicates the benefit of this action outweighs the harms. This recommendation might change when higher quality evidence becomes available.

[Response Ends]

1a.08. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]
Please see information provided in 1a.06 (same question and same response)
[Response Ends]

1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]

Author	Publication	Evidence Grade	Type of Study
Unützer, 2002	Unützer J, Katon W, Callahan CM, et al. Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *JAMA* 2002;288:2836-45.	High	RCT
Trivedi, 2006	Trivedi MH, Fava M, Wisniewski SR, et al. Medication augmentation after the failure of SSRIs for depression. *N Engl J Med* 2006a;354:1243-52.	High	RCT

Author	Publication	Evidence Grade	Type of Study
Hunkeler, 2000	Hunkeler EM, Meresman JF, Hargreaves WA, et al. Efficacy of nurse telehealth care and peer support in augmenting treatment of depression in primary care. *Arch Fam Med* 2000;9:700-08.	High	RCT
Simon, 2000	Simon GE, Van Korff M, Rutter C, Wagner E. Randomised trial of monitoring, feedback, and management of care by telephone to improve treatment of depression in primary care. *BMJ* 2000;320:550-54. (High Quality Evidence)	High	RCT
Trivedi, 2009	Trivedi MH. Tools and strategies for ongoing assessment of depression: a measurement-based approach to remission. *J Clin Psychiatry* 2009;70:26-31.	Low	Observ
Löwe, 2004	Löwe B, Unützer J, Callahan CM, et al. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care* 2004;42:1194-1201.	Low	Cohort
Duffy, 2008	Duffy FF, Chung H, Trivedi M, et al. Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry? *Psychiatric Services* 2008;59:1148-54.	Low	Cohort

Literature Sources, Evidence Grading and Types of Studies

There are several studies with a high quality evidence rating and random control trials evaluated in the systematic review completed by the ICSI guideline work group, but some lower quality observational studies as well, leading to an overall lower quality of evidence rating, but with a strong recommendation for inclusion in clinical practice.

[Response Ends]

1a.10. Provide the estimates of benefit, and consistency across studies.

[Response Begins]
The ICSI guideline workgroup, in its review of all available literature, determined that there was benefit in the ongoing follow-up with patients with major depression and recommend the use of the PHQ-9 tool for both monitoring and the management of depression symptoms.
Recommendation for Collaborative Care:
Benefit-Harms Assessment: Collaborative care has shown to improve patient outcomes and provider satisfaction while decreasing cost outweighing the challenges of implementing a collaborative care program.
Recommendation for Establishing a Follow-Up Plan:
Benefit: Appropriate, reliable follow-up is highly correlated with improved response and remission scores. It is also correlated with the improved safety and efficacy of medications and helps prevent relapse. Harm: Potential harms may include added expense and unnecessary visits. Benefit-Harms Assessment: Benefits appear to outweigh potential harms by a wide margin.

[Response Ends]

1a.11. Indicate what, if any, harms were identified in the study.

[Response Begins]
Please see the guideline workgroup's assessment of benefits and harms in question 1a.10
[Response Ends]

1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]
2021 Submission
The effectiveness of the collaborative care model for depression has also been demonstrated in the adolescent population. Richardson et al. (2014) conducted a randomized controlled trial to examine the collaborative care model vs

usual care for treating adolescents with depression. Results demonstrated that adolescents treated with a collaborative care intervention vs usual care had greater improvement in depressive symptoms at 12 months. The PHQ-9 tool was used to assess the outcome of remission for this study. Using this outcome, the study found that adolescents treated with the collaborative care intervention were significantly more likely to achieve depression remission at both 6 months (OR = 5.2, 95% CI, 1.6-17.3; P = .007) and 12 months (OR = 3.9, 95% CI, 1.5-10.6; P = .007).
Richardson, Laura P., Evette Ludman, Elizabeth McCauley, Jeff Lindenbaum, Cindy Larison, Chuan Zhou, Greg Clarke, David Brent, and Wayne Katon. "Collaborative care for adolescents with depression in primary care: a randomized clinical trial." Jama 312, no. 8 (2014): 809-816.
This newer study is supportive of the collaborative care model and the inclusion of the PHQ-9 for assessing depression symptoms.

[Response Ends]

Group 2 - Evidence - Systematic Reviews Table

1a.03. Provide the title, author, date, citation (including page number) and URL for the systematic review.

[Response Begins]
Guidelines for Adolescents
Guidelines for Adolescent Depression in Primary Care (GLAD-PC): Part I. Practice Preparation, Identification, Assessment, and Initial Management Rachel A. Zuckerbrot, Amy Cheung, Peter S. Jensen, Ruth E.K. Stein, Danielle Laraque and GLAD-PC STEERING GROUP Pediatrics March 2018, 141 (3) e20174081; DOI: https://doi.org/10.1542/peds.2017-4081
https://pediatrics.aappublications.org/content/141/3/e20174081
Guidelines for Adolescent Depression in Primary Care (GLAD-PC): Part II. Treatment and Ongoing Management Amy H. Cheung, Rachel A. Zuckerbrot, Peter S. Jensen, Danielle Laraque, Ruth E.K. Stein and GLAD-PC STEERING GROUP Pediatrics March 2018, 141 (3) e20174082; DOI: https://doi.org/10.1542/peds.2017-4082
https://pediatrics.aappublications.org/content/141/3/e20174082

[Response Ends]

1a.04. Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the systematic review.

[Response Begins]
Identification and Surveillance Recommendation 2: Patients with depression risk factors (e.g., a history of previous depressive episodes, a family history, other psychiatric disorders, substance use, trauma, psychosocial adversity, frequent somatic complaints, previous high-scoring screens without a depression diagnosis, etc.) should be identified (grade of evidence: 2; strength of recommendation: very strong) and systematically monitored over time for the development of a depressive disorder by using a formal depression instrument or tool (targeted screening) (grade of evidence: 2; strength of recommendation: very strong).
Treatment Recommendation 1: PC clinicians should work with administration to organize their clinical settings to reflect best practices in integrated and/or collaborative care models (e.g., facilitating contact with psychiatrists, case managers, embedded therapists). (grade of evidence: 4; strength of recommendation: very strong).
Ongoing Management Recommendation 1: Systematic and regular tracking of goals and outcomes from treatment should be performed, including assessment of depressive symptoms and functioning in several key domains. These include home, school, and peer settings (grade of evidence: 4; strength of recommendation: very strong).

[Response Ends]

1a.05. Provide the grade assigned to the evidence associated with the recommendation, and include the definition of the grade.

[Response Begins]
Grading for each recommendation included in question 1a.04
[Response Ends]

1a.06. Provide all other grades and definitions from the evidence grading system.

[Response Begins]

The level of supporting evidence for each recommendation is based on the Oxford Centre for Evidence-Based Medicine grades of evidence 1–5 system, with 1 to 5 corresponding to strongest to weakest evidence (see http://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf/).

Recommendation strength based on expert consensus was rated in 4 categories: very strong (>90% agreement), strong (>70% agreement), fair (>50% agreement), and weak (<50% agreement). The recommendations in the guidelines were developed only in areas of management that had at least a "strong agreement" among experts.

The original GLAD-PC recommendations were developed on the basis of a synthesis of expert consensus– and evidence-based research review methodologies, as described in Zuckerbrot et al.[17] The 5-step process included conducting focus groups with PC clinicians, patients, and their families, a systematic literature review, a survey of depression experts to address questions that were not answered in the empirical literature,[26] an expert consensus workshop, and an iterative guideline drafting process with opportunity for input from all workshop attendees.

For the research update of the GLAD-PC, systematic literature reviews were conducted in the same 5 key areas of adolescent depression management in PC settings as the original guidelines: identification and assessment, initial management, safety planning, treatment, and ongoing management of youth depression. Consistent with the original review, the updated searches were conducted by using relevant databases (e.g., Medline and PsycInfo), and all primary studies published since the original GLAD-PC reviews in 2005 and 2006 were examined. All update procedures were conducted with the input and guidance of the steering group, which is composed of clinical and research experts, organizational liaisons, and youth and families with lived experience. As in the original review, recommendations were graded on the basis of the University of Oxford's Centre for Evidence-Based Medicine grade of evidence (1–5) system, with 1 to 5 corresponding to the strongest to the weakest evidence respectively (see http://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf). They were also rated on the basis of the strength of expert consensus among the steering group members that the recommended practice is appropriate. Recommendations with strong (>70%) or very strong (>90%) agreement are given here.

Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or n-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, n-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or n-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect	*	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	*	Mechanism-based reasoning

* Cell intentionally left empty
* Level may be graded down on the basis of study quality, imprecision, indirectness (study PICO does not match questions PICO), because of inconsistency between studies, or because the absolute effect size is very small; Level may be graded up if there is a large or very large effect size.
** As always, a systematic review is generally better than an individual study.
How to cite the Levels of Evidence Table
OCEBM Levels of Evidence Working Group*. "The Oxford 2011 Levels of Evidence". Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653
http://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf

[Response Ends]

1a.07. Provide the grade assigned to the recommendation, with definition of the grade.

[Response Begins]
Grading for each recommendation included in question 1a.04
[Response Ends]

1a.08. Provide all other grades and definitions from the recommendation grading system.

[Response Begins]
Please refer to responses in question 1a.06
[Response Ends]

1a.09. Detail the quantity (how many studies) and quality (the type of studies) of the evidence.

[Response Begins]
The updated GLAD PC guideline includes both screening and ongoing assessment for depression symptoms and outcomes. While the original guideline (2007) focused on a suggested tool, the modified PHQ-9 for adolescents which was created for use within the guideline, the updated guideline explores a variety of tools which include the PHQ-9 (below, see guideline for reference list https://pediatrics.aappublications.org/content/141/3/e20174081#sec-6)
Most relevant were the 2 publications by Richardson et al[56,57] in which they validated the Patient Health Questionnaire-2 (PHQ-2) and the Patient Health Questionnaire-9 (PHQ-9) in a PC sample against a gold standard diagnostic interview (the Diagnostic Interview Schedule for Children-IV [DISC-IV]).
Researchers have looked at brief depression-specific screening questions that stand alone (e.g., the PHQ-2),[51,57,65,75,79,82,85] longer depression-specific scales that stand alone (e.g., the PHQ-9, the Mood and Feelings Questionnaire, the Columbia Depression Scale, and the PHQ-9: Modified for Teens),[58,62,63,66,67,70,74,78,80–82,86–88] brief depression screening questions that are part of a larger psychosocial tool (e.g., the Guidelines for Adolescent Preventive Services [GAPS] questionnaire and the Pediatric Symptom Checklist [PSC]),[53,54,64,68,69] and brief screening questions or longer depression-specific scales that are combined with other screens for either other psychiatric disorders (e.g., Screen for Child Anxiety Related Disorders-5) and/or screens for other high-risk behaviors (e.g., substance use and sexual activity) to make a more multidimensional tool or packet in 1 (e.g., the behavioral health screen [BHS]).[50,52,55,59–61,76,77,83,84,89] Not all of the screens in these studies have specific psychometric validation data (e.g., 2 depression questions on the GAPS). Clinicians may also consider the use of tools that can be used to screen for depression and other risk behaviors or more disorders. Although no researchers have compared the functional or depressive outcomes of a cohort of adolescents who were initially screened only for depression with a cohort of adolescents who were initially screened for an array of high-risk behaviors and

emotional issues, some hint at the possibility that too much information may overwhelm the clinician and result in positive depression screening questions being overlooked in the morass of issues needing to be addressed.[52,53,59-61,64,76,80,82-84,89] Therefore, clinicians should base the selection of a depression-specific tool versus a more general tool on their own expertise and clinical supports in their practices. For example, a solo practitioner starting to address depression care in his or her practice may choose to start with screening for depression alone before moving to more general screening for riskier behaviors or disorders.

[Response Ends]

1a.10. Provide the estimates of benefit, and consistency across studies.

[Response Begins]
not available
[Response Ends]

1a.11. Indicate what, if any, harms were identified in the study.

[Response Begins]
not available
[Response Ends]

1a.12. Identify any new studies conducted since the systematic review, and indicate whether the new studies change the conclusions from the systematic review.

[Response Begins]
not applicable
[Response Ends]

1a.13. If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, describe the evidence on which you are basing the performance measure.

[Response Begins]
not applicable
[Response Ends]

1a.14. Briefly synthesize the evidence that supports the measure.

[Response Begins]
not applicable
[Response Ends]

1a.15. Detail the process used to identify the evidence.

[Response Begins]
not applicable
[Response Ends]

1a.16. Provide the citation(s) for the evidence.

[Response Begins]
not applicable
[Response Ends]

1b.01. Briefly explain the rationale for this measure.

Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.

[Response Begins]

Adults:

Depression is a common and treatable mental disorder. The Centers for Disease Control and Prevention states that an estimated 6.6% of the U.S. adult population (14.8 million people) experiences major depressive disorder during any given 12-month period. Additionally, dysthymia accounts for an additional 3.3 million Americans. In 2006 and 2008, an estimated 9.1% of U.S. adults reported symptoms for current depression.1 Persons with a current diagnosis of depression and a lifetime diagnosis of depression or anxiety were significantly more likely than persons without these conditions to have cardiovascular disease, diabetes, asthma and obesity and to be a current smoker, to be physically inactive and to drink heavily.2 People who suffer from depression have lower incomes, lower educational attainment and fewer days working days each year, leading to seven fewer weeks of work per year, a loss of 20% in potential income and a lifetime loss for each family who has a depressed family member of $300,000.3 The cost of depression (lost productivity and increased medical expense) in the United States is $83 billion each year.4

Prevalence updates: 2019 National Survey on Drug Use and Health (NSDUH) estimates that 19.4 million adults or 7.8% had at least one major depressive episode with the highest prevalence of 15.2% among individuals aged 18 - 25.14

Adolescents and Adults:

The Centers for Disease Control and Prevention states that during 2009-2012 an estimated 7.6% of the U.S. population aged 12 and over had depression, including 3% of Americans with severe depressive symptoms. Almost 43% of persons with severe depressive symptoms reported serious difficulties in work, home and social activities, yet only 35% reported having contact with a mental health professional in the past year.5

Depression is associated with higher mortality rates in all age groups. People who are depressed are 30 times more likely to take their own lives than people who are not depressed and five times more likely to abuse drugs.6 Depression is the leading cause of medical disability for people aged 14 – 44.7 Depressed people lose 5.6 hours of productive work every week when they are depressed, fifty percent of which is due to absenteeism and short-term disability.

Adolescents:

In 2014, an estimated 2.8 million adolescents age 12 to 17 in the United States had at least one major depressive episode in the past year. This represented 11.4% of the U.S. population. The same survey found that only 41.2 percent of those who had a Major Depressive Episode received treatment in the past year.8 The 2013 Youth Risk Behavior Survey of students grades 9 to 12 indicated that during the past 12 months 39.1% (F) and 20.8% (M) indicated feeling sad or hopeless almost every day for at least 2 weeks, planned suicide attempt 16.9% (F) and 10.3% (M), with attempted suicide 10.6% (F) and 5.4% (M).9 Adolescent-onset depression is associated with chronic depression in adulthood.10 Many mental health conditions (anxiety, bipolar, depression, eating disorders, and substance abuse) are evident by age 14.11 The 12-month prevalence of MDEs increased from 8.7% in 2005 to 11.3% in 2014 in adolescents and from 8.8% to 9.6% in young adults (both $P < .001$). The increase was larger and statistically significant only in the age range of 12 to 20 years. The trends remained significant after adjustment for substance use disorders and sociodemographic factors. Mental health care contacts overall did not change over time; however, the use of specialty mental health providers increased in adolescents and young adults, and the use of prescription medications and inpatient hospitalizations increased in adolescents. 12 In 2015, 9.7% of adolescents in MN who were screened for depression or other mental health conditions, screened positively.13

References

1. CDC. Current Depression Among Adults --- United States, 2006 and 2008. MMWR 2010;59(38);1229-1235.
2. Strine TW, Mokdad AH, Balluz LS, et al. Depression and anxiety in the United States: findings from the 2006 Behavioral Risk Factor Surveillance System. Psychiatr Serv 2008;59:1383--90.
3. Smith, J. P., & Smith, G. C. (2010). Long-term economic costs of psychological problems during childhood. Social Science & Medicine, 71, 110-115.
4. Greenberg, P. E., Kessler, R. C., Birnbaum, H. G., Leong, S. A., Lowe, S. W., Berglund, P. A., et al. (2003). The economic burden of depression in the United States: How did it change between 1990 and 2000? Journal of Clinical Psychiatry, 64, 1465-1475.
5. Pratt LA, Brody DJ. Depression in the U.S. household population, 2009–2012. NCHS data brief, no 172. Hyattsville, MD: National Center for Health Statistics. 2014.
6. Joiner, Thomas Myths about suicide. Cambridge, MA, US: Harvard University Press. (2010). 288 pp.7. Stewart, W. F., Ricci, J. A., Chee, E., Hahn, S. R., & Morganstein, D. (2003). Cost of lost productive work time among US workers with depression. Journal of the American Medical Association, 289, 3135-3144.

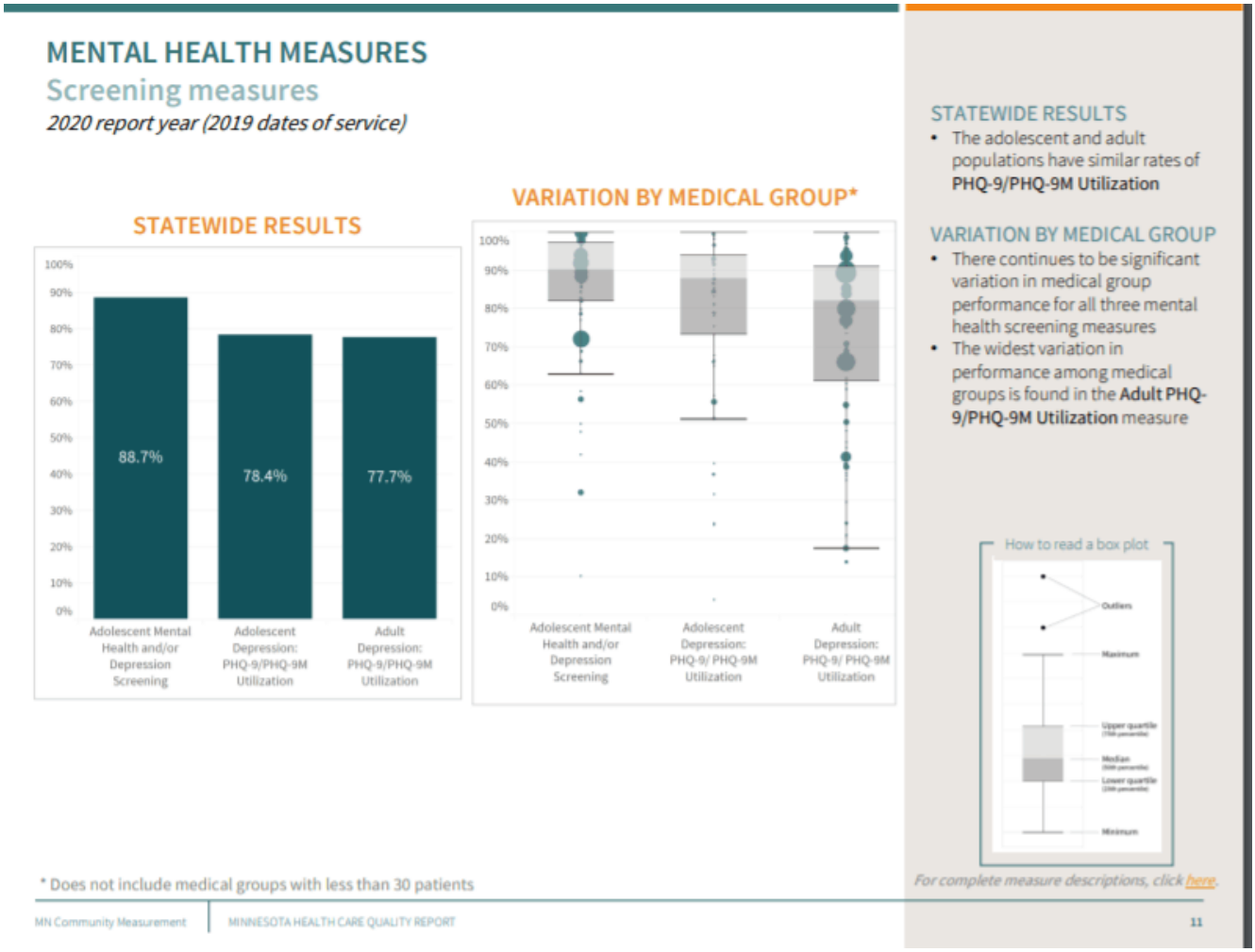
7. National Institute Mental Health/ National Institute Health 2014 prevalence of depression in adolescents statistics www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adolescents.shtml
8. 2013 Youth Risk Behavior Survey Suicide and suicide Attempts in Adolescents Clinical Report American Academy of Pediatrics July 2016
9. Lewinsohn, P. M., Rohde, P., Klein, D. N., & Seeley, J. R. (1999). Natural course of adolescent major depressive disorder: I. Continuity into young adulthood. Journal of the American Academy of Child & Adolescent Psychiatry, 38(1), 56-63.
10. Why do many psychiatric disorders emerge during adolescence? Giedd et al. Nat Rev Neurosci
1. Dec 2008 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2762785/
11. National Trends in the Prevalence and Treatment of Depression in Adolescents and Young Adults. Ramin, M et al Pediatrics November 2016 http://pediatrics.aappublications.org/content/early/2016/11/10/peds.2016-1878
12. New Measures Evaluate Rates of Obesity Counseling for Kids, Depression Screening for Teens Oct 2015 www.mncm.org/new-measures-evaluate-rates-of-obesity-counseling-for-kids-depression-screening-for-teens/
13. National Institutes of Health Transforming the Understanding and Treatment of Mental Illness https://www.nimh.nih.gov/health/statistics/major-depression

[Response Ends]

1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.

Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]
Minnesota Statewide Reporting
Average: 77.7% for Adults (n = 248,163) and 78.4% for Adolescents (n = 19,574). Variability among medical groups is displayed by the range of results (25 to 100% and 8 to 100% respectively. Box plot diagrams further display the wide variability in medical group rates; for the adults, a significant portion of clinics are in the lower quartile.



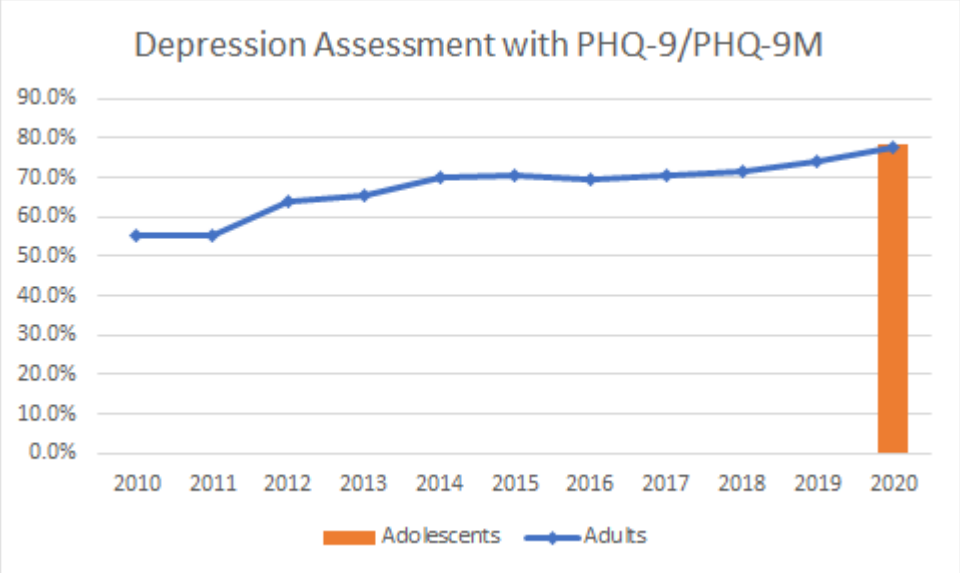
MNCM Statewide Reporting for Mental Health Measures; Health Care Quality Report 2020

The image above depicts the variability of rates among medical groups around the statewide average:

- 77.7% for Adults (n = 248,163)
- 78.4% for Adolescents (n = 19,574).

Variability among medical groups is displayed by the range of results (25 to 100% and 8 to 100% respectively. Box plot diagrams further display the wide variability in medical group rates; for the adults, a significant portion of clinics are in the lower quartile.

Rates for this measure over time. Adolescents were incorporated into the measure in report year 2020.



Trend of Rates of Depression Assessment; Adolescents added report year 2020

Trend of rates over the past ten years for adults demonstrate gradual improvement from 55.0% to 77.7%

[Response Ends]

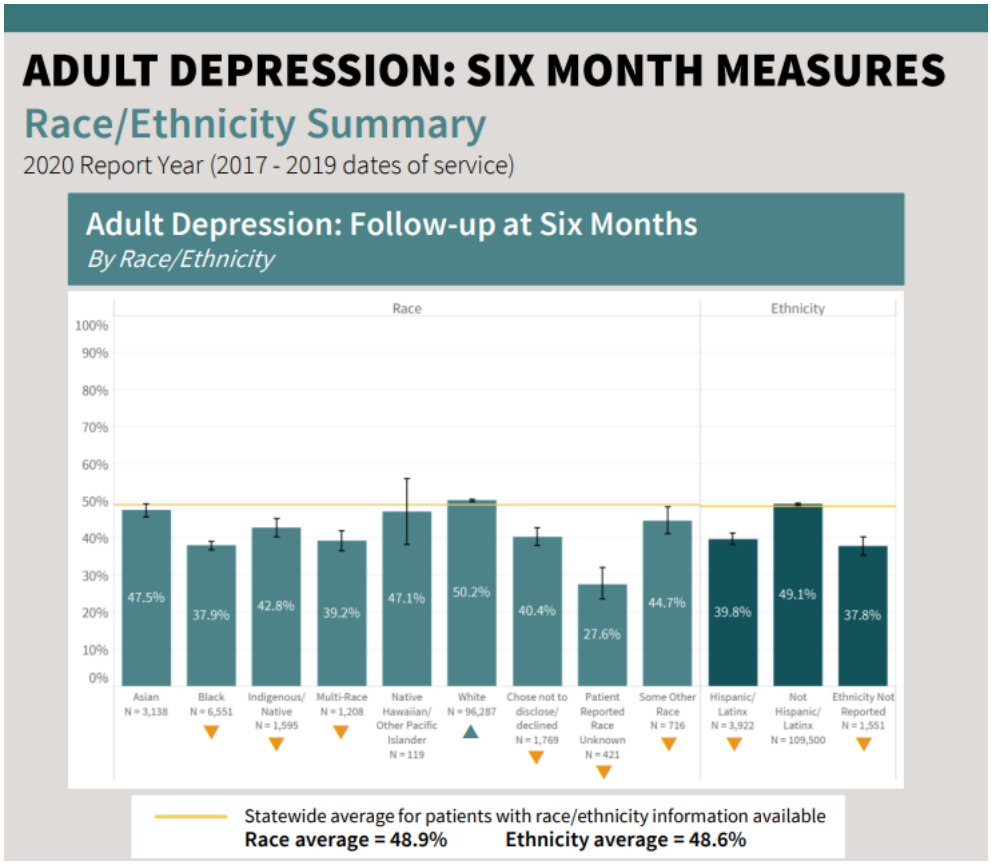
1b.03. If no or limited performance data on the measure as specified is reported above, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement. Include citations.

[Response Begins]
not applicable
[Response Ends]

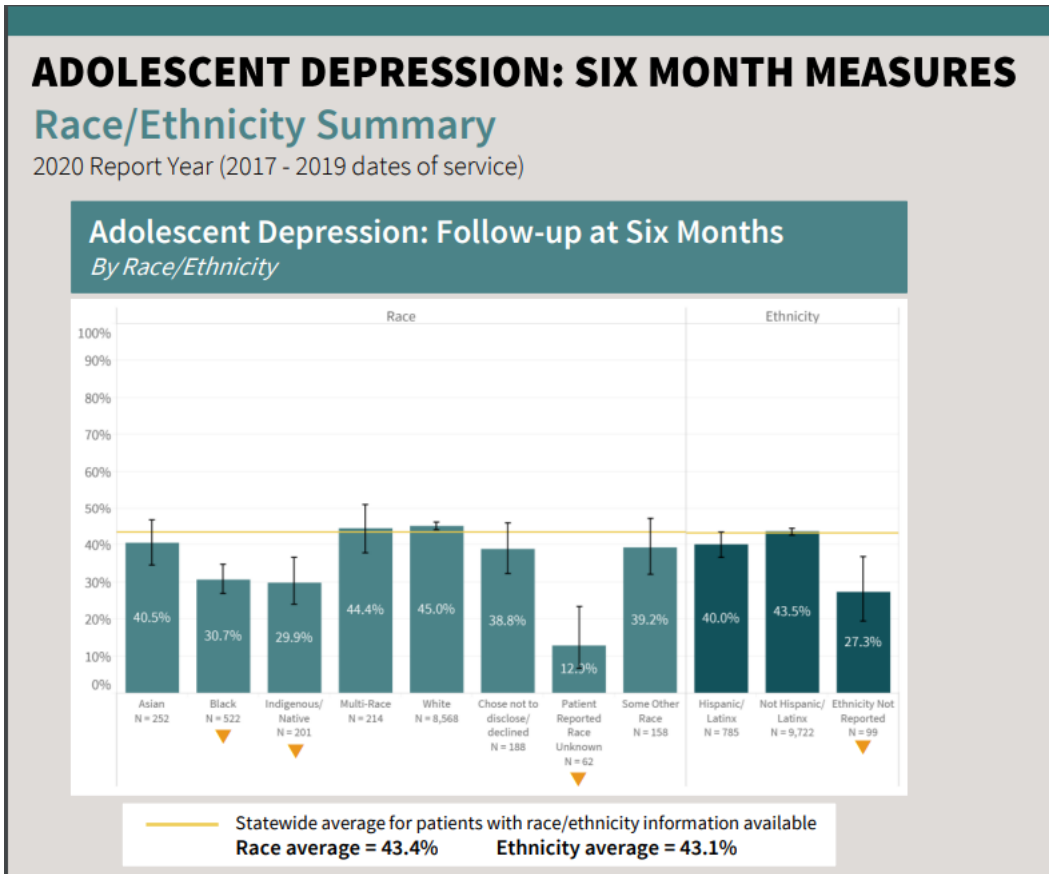
1b.04. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]
This process, PRO based measure is a companion measure to support outcome measures of remission and response at six and twelve months (NQF# 0710, 0711, 1884 and 1885) and our Health Care Disparities Reports publicly report outcome and follow-up rates by
Annual Health Care Disparities Report
Publicly available at https://mncm.org/reports/#community-reports
https://mncm.org/reports/#community-report



Display of Related Outcome Measures at Six Months by Race and Ethnicity; Adults
Adults who are Black, Indigenous/Native, Multi-Race or Hispanic/Latinx are among those who have significantly lower rates of depression follow-up, response and remission at six months compared to the race/ethnicity averages. Additionally, adults who are Asian have significantly lower rates of depression response and remission at six months.



Display of Related Outcome Measures at Six Months by Race and Ethnicity; Adolescents
Adolescents who are Black have significantly lower rates of follow-up, response and remission at six months compared to the race/ethnicity averages. Patients who are Indigenous/Native have significantly lower rates of followup at six months compared to the race average.

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, 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 above.

[Response Begins]
Source: ICSI Guideline for Major Depression in Adults in Primary Care 17th edition March 2016
https://www.icsi.org/_asset/fnhdm3/Depr.pdf
Many patients with major depression do not initially complain of depressed mood or anehdonia (American Psychiatric Association, 2013). Clinicians need to suspect this diagnosis based on a profile of common presentations and risk factors, taking into account cultural considerations (American Psychiatric Association, 2013).
Clinicians should acknowledge the impact of culture and cultural differences on physical and mental health. There is evidence that non-majority racial and cultural groups in the U.S. are less likely to be treated for depression than European Americans. In an epidemiological study that compared rates of diagnosing and treating depression in the early 1990s to patterns 10 years later, only 4.9% of minorities were treated with antidepressants compared with 12.4% of non-Hispanic Caucasians (Mojtabai, 2008).

A person´s cultural and personal experiences influence his/her beliefs and therefore attitudes and preferences. If these experiences are taken into consideration, openness to and readiness to change (including readiness to seek and adhere to treatment) will be enhanced. People of differing racial/ethnic groups are optimally treated using currently available evidence-based interventions when differential personal elements, from biological to environmental to cultural, are considered during the treatment planning process (Schraufnagel, 2006).
Cultural beliefs and common presentations

- When dealing with patients from diverse cultures, the impact of patient´s cultural beliefs around depression, cultural stigma and manifestation of depression in physical symptoms vs. psychological can play a role in how patients perceive depression and subsequently seek treatment (Kleinman, 2004).
- Clinicians can create a more comfortable environment for a patient of another culture by acknowledging the impact of culture and cultural differences on physical and mental health (Muñoz, 2005; Miranda, 2004).
- Bodily idioms of distress are very common in many cultures. In place of psychosomatic theories that emphasize individuals´ inner conflict, many traditions of medicine have sociosomatic theories that link bodily and emotional distress to problems in the social world (Kirmayer, 2001).
- The concept of depression varies across cultures. For example, in many cultures, for depression to become a problem for which a person seeks medical treatment, symptoms may include psychosis, conversion disorders or significant physical ailments (Karasz, 2005).

Age disparities have also been documented. Depression in the elderly is widespread, often undiagnosed and usually untreated. It is a common misperception that it is a part of normal aging. Losses, social isolation and chronic medical problems that older patients experience can contribute to depression. The rate of depression in adults older than 65 years of age ranges from 17% to 37% in primary care settings and is between 14 and 42% in the elderly who live in long term care facilities. Among adolescents, some estimates suggest that only 25 percent of adolescents diagnosed with depression receive treatment; among those who go undetected, 20 percent develop recurrent or chronic depression (O'Connor, 2009; Garber, 2009).

[Response Ends]

Criteria 2: 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.

spma.01. Indicate whether there are changes to the specifications since the last updates/submission. If yes, update the specifications in the Measure Specifications section of the Measure Submission Form, and explain your reasoning for the changes below.

[Response Begins]
Yes
[Yes Please Explain]
Several changes to the measure specifications were made:

- incorporating adolescents ages 12 to 17
- added PHQ-9M (modified for teens) PRO tool
- modified exclusion value set for personality disorder
- added exclusions for schizophrenia and pervasive developmental disorder

[Response Ends]

spma.02. Briefly describe any important changes to the measure specifications since the last measure update and provide a rationale.

For annual updates, please explain how the change in specifications affects the measure results. If a material change in specification is identified, data from re-testing of the measure with the new specifications is required for early maintenance review.

For example, specifications may have been updated based on suggestions from a previous NQF CDP review.

[Response Begins]
Since the last maintenance update, we convened our multi-stakeholder expert workgroup to consider modifying the measure to include adolescents as well as reviewing related measure construct components. As a result of our process, we are updating the measures to add the adolescent population; add the PHQ-9M tool; tighten up the personality disorders exclusions list; add exclusions for schizophrenia and pervasive developmental disorders and simplify the diagnosis criterion. Details are as follows:
For 2020 Report Year (dates of index event 1/1/2018 to 12/31/2018)
1. Incorporate adolescents into the depression measures
* Modify age range to include adolescents; age 12 and older
* Report measures as two separate stratifications by age (not combined); ages 12 to 17 and ages 18 and older
Reason: The U.S. Preventive Services Task Force and other guideline organizations recommend screening adolescents for depression. Depression is a significant problem for adolescents, affecting an estimated 11% of the population. Many mental health conditions are evident by age 14 and the consequences of adolescent depression can have a lifelong impact.
2. Patient Reported Outcome Tools for numerator are the PHQ-9 and PHQ-9M
* Add the PHQ-9M as a PRO tool that can be used
* Providers may elect to use either tool; no measure construct restriction for age. For example, if a family practice clinic is currently using the PHQ-9 tool for their adult patients, they can elect to use the same tool for ages 12 to 17. Likewise, if a pediatric clinic is using the PHQ-9M in their practice, they can decide to administer the PHQ-9M to their 18/19/20 year old patients.
Reason: The expert panel reviewed 21 additional tools against standardized criteria and concluded very few had cut-points for severity levels of depression or remission. Further, using PRO tools with significantly different numbers of questions could impact the response measures (50% or greater in improvement of scores) in addition to adversely affecting denominator comparability. For example, if one practice is using the Beck BDI-II tool (21 questions/ total score 63/ denominator > 19/ remission < 14) and another practice is using the PHQ-9 (9 questions/ total score 27/ denominator

> 9/ remission < 5), it can't be assured that the two tools are identifying the denominator of patients in the exact same way.
3. Modifications to exclusions include the following:
* Personality disorders narrowed to emotionally labile conditions and moved to the allowable exclusion category
* Add exclusion value set for schizophrenia or psychotic disorder as a required exclusion
* Add exclusion value set for pervasive developmental disorder as an allowable exclusion
Reason: The expert panel determined these conditions may require a different course of treatment, and holding a provider responsible for remission/response within the timeframe defined by the measure may be inappropriate. In addition, the NQF Behavioral Steering Committee requested we examine the personality disorder exclusion.
4. For behavioral health settings, remove the requirement that the diagnosis of major depression or dysthymia must be in the primary position.
* Relates to new exclusion for schizophrenia or psychotic disorder; no longer necessary
Reason: simplification of measures, position order of diagnosis is irrelevant in behavioral health settings.

[Response Ends]

sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see What Good Looks Like).

[Response Begins]
Depression Assessment with PHQ-9/ PHQ-9M
[Response Ends]

sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

[Response Begins]
The percentage of adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) with a diagnosis of major depression or dysthymia who have a completed PHQ-9 or PHQ-9M tool during a four month measurement period.
[Response Ends]

sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Surgery: General*

[Response Begins]
Behavioral Health: Depression
[Response Ends]

sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

[Response Begins]
Health and Functional Status
[Response Ends]

sp.06. Select one or more target population categories.

Select only those target populations which can be stratified in the reporting of the measure's result.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]
Adults (Age >= 18)
Children (Age < 18)
Elderly (Age >= 65)
[Response Ends]

sp.07. Select the levels of analysis that apply to your measure.

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Clinician: Group/Practice
[Response Ends]

sp.08. Indicate the care settings that apply to your measure.

Check ONLY the settings for which the measure is SPECIFIED and TESTED.
[Response Begins]
Outpatient Services
[Response Ends]

sp.09. 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. If no URL is available, indicate "none available".

[Response Begins]
https://helpdesk.mncm.org/helpdesk/KB/View/24186732-data-collection-technical-guide--depression-care
[Response Ends]

sp.11. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, contact staff. Provide descriptors for any codes. Use one file with multiple worksheets, if needed.
[Response Begins]
Available in attached Excel or csv file
[Response Ends]

Attachment: MNCM Depression Care VS Specs Definitions w Redesign 6-9-2021.xlsx

sp.12. State the numerator.

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.

[Response Begins]
Adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) included in the denominator who have at least one PHQ-9 or PHQ-9M tool administered and completed during a four month measurement period.
[Response Ends]

sp.13. Provide details needed to calculate the numerator.

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 sp.11.

[Response Begins]
The total number of unique adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) in the denominator who had a least one PHQ-9 or PHQ-9M tool administered and completed during a four month measurement period in which a visit or contact with the patient has occurred.
Partially completed tools (e.g. answering 6 of the 9 questions) do not count as a completed tool. A valid PHQ-9 or PHQ-9M requires the completion of all nine questions for accurate scoring.
The numerator rate is calculated as follows:
pts with major depression or dysthymia with one or more completed PHQ-9 or PHQ-9M tools/
pts with major depression or dysthymia with a visit or contact during the measurement period
Rates are stratified by adolescents (12 to 17 years of age) and adults (18 years of age or older).
Time period for data collection: four month measurement periods (In the MN program 2/01 to 5/31, 6/01 to 9/30 and 10/01 to 1/31) with dates of service occurring within the four month period.

[Response Ends]

sp.14. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]
Adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) with a diagnosis of major depression or dysthymia.
[Response Ends]

sp.15. Provide details needed to calculate the denominator.

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 sp.11.

[Response Begins]
The target population, patients age 12 and older with the diagnosis of major depression or dysthymia, regardless of severity level of the PHQ-9 or PHQ-9M.

The number of unique patients who had a least one visit or contact with a provider during the measurement period with a diagnosis of major depression or dysthymia (Major Depression or Dysthymia Value Set). Contact is defined as visit, telephone call, e-visit or other contact that is associated with a PHQ-9 tool being completed by the patient.

[Response Ends]

sp.16. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]
Patients who die, are a permanent resident of a nursing home or are enrolled in hospice are excluded from this measure. Additionally, patients who have a diagnosis of bipolar or personality disorder, schizophrenia or psychotic disorder, or pervasive developmental disorder are excluded.
[Response Ends]

sp.17. Provide details needed to calculate the denominator exclusions.

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 sp.11.

[Response Begins]
Required exclusions:

- Patient had a diagnosis of Bipolar Disorder (*Bipolar Disorder* Value Set) any time prior to the end of the measurement period
- Patient had an active diagnosis of Schizophrenia or Psychotic Disorder (*Schizophrenia Psychotic Disorder* Value Set) any time prior to the end of the measurement period

Allowable exclusions:

- Patient died prior to the end of the measurement period
- Patient was a permanent nursing home resident at any time during the measurement period
- Patient was in hospice or receiving palliative care at any time during the measurement period (*Palliative Care* Value Set)
- Patient had a diagnosis of Personality Disorder – Emotionally Labile Conditions (*Personality Disorder – Emotionally Labile* Value Set) any time prior to the end of the measurement period
- Patient had an active diagnosis of Pervasive Developmental Disorder (*Pervasive Disorder* Value Set) any time prior to the end of the measurement period
- The direct data submission process in MN allows for both up-front exclusions of the population and, because this is a longitudinal outcome measure, processes are in place to allow exclusions that may occur after index during the course of the measurement assessment period. Please see field specifications in the attached data dictionary.

[Response Ends]

sp.18. Provide all information required to stratify the measure results, if necessary.

Include 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 in the Data Dictionary field.

[Response Begins]
This measure is stratified by age range and results are reported separately by age: Adolescents (12-17 years of age) and Adults (18 years of age and older).
[Response Ends]

sp.19. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.
[Response Begins]
No risk adjustment or risk stratification
[Response Ends]

sp.20. Select the most relevant type of score.

Attachment: If available, please provide a sample report.
[Response Begins]
Rate/proportion
[Response Ends]

sp.21. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score
[Response Begins]
Better quality = Higher score
[Response Ends]

sp.22. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.

[Response Begins]
Historically, this measure is calculated by submitting a count of patients for the denominator and a count of patients in the numerator to a HIPAA secure data portal as part of the process in uploading a detailed patient file to calculate the six and twelve month remission outcome rates. MNCM is in the process of onboarding MN practice to a new warehouse (PIPE) and will calculate this measure centrally for practices based on encounter level data; full statewide transition to PIPE is planned for 2024.
The numerator rate is calculated as follows:
# of adolescent and adult pts with major depression or dysthymia with at least one PHQ-9 or PHQ-9M tool administered during the four month measurement period/
# of adolescent and adult pts with major depression or dysthymia
Query processes that medical groups follow to obtain counts:
During the four month measurement period (e.g. dates of service 6/1/2020 to 9/30/2020) how many patients had an office visit or other contact (phone, email) and diagnosis codes for major depression or dysthymia? (denominator)
Of these patients, how many had a PHQ-9 or PHQ-9M tool administered? (numerator)
The counting process is validated during the denominator certification process (where groups document all steps in identifying the depression population). Groups are asked to describe the process they use for obtaining the counts. Denominator documents are reviewed (certified) by MNCM staff prior to data collection and submission. This is to insure that all groups are identifying their population correctly.

[Response Ends]

sp.23. Attach a copy of the instrument (e.g. survey, tool, questionnaire, scale) used as a data source for your measure, if available.

[Response Begins]
Copy of instrument is attached.

[Response Ends]

Attachment: 0712_PHQ-9-Modified-For-Teens-64711 GLAD-PC.pdf
Attachment: PHQ9.pdf

sp.24. Indicate the responder for your instrument.

[Response Begins]
Patient
[Response Ends]

sp.25. If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

[Response Begins]
The measure and its denominator are not based on a sample. The measure was developed with the intent of full population reporting with the EMR as the data source.
[Response Ends]

sp.26. Identify whether and how proxy responses are allowed.

[Response Begins]
Proxy responses are not allowed, the PRO tool has to be completed by the patient. The tool is validated for multiple modes of administration and is translated and available in more than 90 languages. https://www.phqscreeners.com/select-screener
[Response Ends]

sp.27. Survey/Patient-reported data.

Provide instructions for data collection and guidance on minimum response rate. Specify calculation of response rates to be reported with performance measure results.

[Response Begins]
PROM Developer Instruction manual: www.phqscreeners.com
PHQ-9 Depression Severity. This is calculated by assigning scores of 0, 1, 2, and 3, to the response categories of "not at all", "several days", "more than half the days", and "nearly every day" respectively. PHQ-9 total score for the nine items ranges from 0 to 27. Scores of 5, 10, 15, and 20 represent cut-points for mild, moderate, moderately severe and severe depression, respectively. Sensitivity to change has also been confirmed.
Use of the tool for measurement: All nine questions need to be completed/ answered for a valid score. Patient responses are not imputed and the tool score is derived from a simple summation of the responses.
The internal reliability of the PHQ-9 was excellent, with a Cronbach´s alpha of 0.89 in the PHQ Primary Care Study and 0.86 in the PHQ Ob-Gyn Study. Test-retest reliability of the PHQ-9 was also excellent. Correlation between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours was 0.84, and the mean scores were nearly identical (5.08 vs 5.03).
PHQ-9 has been validated in adolescent populations (age 13 to 17), as well as adults and elderly.
Kronke K., Spitzer R. The PHQ-9 Validity of a Brief Depression Severity Measure J Gen Intern Med 2001 September; 16(9): 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x PMCID: PMC1495268
Lowe B., Unutzer J. Monitoring Depression Treatment outcomes with the Patient Health Questionnaire-9 Medical Care Volume 42 Number 12 December 2004
Duffy F., Chung H. Systematic Use of Patient-Rated Depression Severity Monitoring: Is It Helpful and Feasible in Clinical Psychiatry? Psychiatric Services October 2008 Vol. 59 No. 10
Richardson L., McCauley E. Evaluation of the Patient Health Questionnaire (PHQ-9) for Detecting Major Depression among Adolescents Pediatrics 2010 December; 126(6): 1117–1123. doi:10.1542/peds.2010-0852.
The PHQ-9M Modified for Teens is the PHQ-9 tool with slight wording adjustment (in CAPS below) in three questions in order to tailor the tool for the adolescent population with age-appropriate terms.
Q2: Feeling down, depressed, IRRITABLE, or hopeless?

Q5: Poor appetite, WEIGHT LOSS, or overeating?
Q7: Trouble concentrating on things like SCHOOL WORK, reading, or watching TV?
Otherwise, the nine questions used in scoring the tool are identical to the PHQ-9.
The copyright statement on the PHQ-9M tool is stated: "Modified with permission by the GLAD-PC team from the PHQ-9 (Spitzer, Williams & Kroenke, 1999), Revised PHQ-A (Johnson, 2002) and the CDS (DISC Development Group, 2000)"
Although widely used in pediatric practices and endorsed by the AAP, APA and AACAP, the modified version of the PHQ-9 tool has not had separate validation studies, as the nine questions are essentially the same as the original PHQ-9, which was been validated for the adolescent population (ages 13 and older). The APA recommends using the modified version of the PHQ-9 for children ages 11 to 17 to assess depression symptom severity (APA, 2015).
American Psychiatric Association. 2015. Online Assessment Measures. Severity Measure for Depression, Child Age 11 to 17 (PHQ-9 modified for Adolescents [PHQ-A], Adapted). https://www.psychiatry.org/psychiatrists/practice/dsm/dsm-5/online-assessment-measures

[Response Ends]

sp.28. Select only the data sources for which the measure is specified.

[Response Begins]
Electronic Health Records
[Response Ends]

sp.29. Identify the specific data source or data collection instrument.

For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.

[Response Begins]
The data source is the medical group's/ clinic's medical record information, most frequently from an EMR. A CSV file is created by each medical group and uploaded to a password protected, HIPAA secure data portal which performs rate calculation.
PROM
The PHQ-9 depression assessment tool is a patient reported outcome tool that is in the public domain and can be obtained for free use on the Patient Health Questionnaire (PHQ) Screeners website at www.phqscreeners.com. Modes of administration include traditional paper, mail, electronic and telephonic. The tool is available on the website with 79 language translations available.
The PHQ-9 tool is validated for use as a measure to assess the level of depression severity (for initial treatment decisions) as well as an outcome tool (to determine treatment response). [Löwe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. Med Care 2004;42:1194-1201 and Kroenke K, Spitzer RL, Williams JBW, Löwe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010]
The PHQ-9M is a modified version of the PHQ-9 tool for adolescents. Please refer to discussion in question sp.27

[Response Ends]

sp.30. Provide the data collection instrument.

[Response Begins]
Available at measure-specific web page URL identified in sp.09
[Response Ends]

2ma.01. Indicate whether additional empirical reliability testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Reliability - Testing. Include information on all testing conducted (prior testing as well as any new testing).

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:
Current Submission:
Updated testing information here.
Previous Submission:
Testing from the previous submission here.

[Response Begins]
Yes
[Response Ends]

2ma.02. Indicate whether additional empirical validity testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Validity - Testing. Include information on all testing conducted (prior testing as well as any new testing).

Please separate added or updated information from the most recent measure evaluation within each question response in the Scientific Acceptability sections. For example:
Current Submission:
Updated testing information here.
Previous Submission:
Testing from the previous submission here.

[Response Begins]
Yes
[Response Ends]

2ma.03. For outcome, patient-reported outcome, resource use, cost, and some process measures, risk adjustment/stratification may be conducted. Did you perform a risk adjustment or stratification analysis?

[Response Begins]
No
[Response Ends]

2ma.04. For maintenance measures in which risk adjustment/stratification has been performed, indicate whether additional risk adjustment testing has been conducted since the most recent maintenance evaluation. This may include updates to the risk adjustment analysis with additional clinical, demographic, and social risk factors.

Please update the Scientific Acceptability: Validity - Other Threats to Validity section.

Note: This section must be updated even if social risk factors are not included in the risk adjustment strategy.

[Response Begins]
No additional risk adjustment analysis included
[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

- Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.
- All required sections must be completed.

- For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.
- An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
- Contact NQF staff with any questions. Check for resources at the Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for the 2021 Measure Evaluation Criteria and Guidance.

Note: The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.
2a. Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For instrument-based measures (including PRO-PMs) and composite performance measures, reliability should be demonstrated for the computed performance score.
2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.
2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure;
AND
If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).
2b3. For outcome measures and other measures when indicated (e.g., resource use):

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

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful 16 differences in performance;
OR
there is evidence of overall less-than-optimal performance.
2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.
2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(if not conducted or results not adequate, justification must be submitted and accepted)

Definitions
Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of

the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.
Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
Risk factors that influence outcomes should not be specified as exclusions.
With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v.$5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Scientific Acceptability sections. For example:
2021 Submission:
Updated testing information here.
2018 Submission:
Testing from the previous submission here.

2a.01. Select only the data sources for which the measure is tested.

[Response Begins]
Electronic Health Records
[Response Ends]

2a.02. 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).

[Response Begins]
This measure is in full implementation with submission of data from all primary care and behavioral health (psychiatry) clinics in Minnesota. For this measure, due to the condition's chronic episodic nature, no sampling is allowed and the full population of eligible patients, regardless of payer, is included.
Please note that the data source is electronic health record; all primary care and behavioral health clinics in MN are on electronic health records, therefore the data source for testing no longer includes paper records.

[Response Ends]

2a.03. Provide the dates of the data used in testing.

Use the following format: "MM-DD-YYYY - MM-DD-YYYY"

[Response Begins]
2021 Submission
Adult patients with dates of service 10/1/2019 to 1/31/2020 reported in 2020, Adolescent patients age 12 to 17 with dates of service 1/1/2019 to 12/31/2020.

[Response Ends]

2a.04. Select the levels of analysis for which the measure is tested.

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.
Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.
Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]
Clinician: Group/Practice
[Response Ends]

2a.05. List the measured entities 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.

[Response Begins]
2021 Submission
Sites represent all primary care and behavioral health (psychiatry) clinics in Minnesota. Clinics that are in bordering cities in other states (Wisconsin, North Dakota) that wish to participate as part of their health system (also located in MN) may do so, but are not required to participate. Clinics represent urban and rural, large multi-specialty health care systems, medium and small practices that care for adult patients with depression. 103 medical groups representing 615 clinics were included in the testing of this measure, representing 227, 127 adults and 12,616 adolescents.

[Response Ends]

2a.06. Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.

If there is a minimum case count used for testing, that minimum must be reflected in the specifications.

[Response Begins]
2021 Submission
227,127 adult patients and 12,616 adolescents were included for testing and analysis. There was no elimination of patients based on age, race/ethnicity, or diagnosis with the exception of valid clinical co-morbid diagnoses for exclusions (bi-polar disorder and personality disorder) which are already excluded from the denominator.

[Response Ends]

2a.07. 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.

[Response Begins]
Reliability and validity statistics performed at the clinic level for all clinics with $\geq$ 30 patients in the denominator.
[Response Ends]

2a.08. List 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.

[Response Begins]
This process measure is not risk adjusted, therefore social risk factors were not assessed for this measure.
[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.07 check patient or encounter-level data; in 2a.08 enter "see validity testing section of data elements"; and enter "N/A" for 2a.09 and 2a.10.

2a.09. Select the level of reliability testing conducted.

Choose one or both levels.
[Response Begins]
Accountable Entity Level (e.g., signal-to-noise analysis)
[Response Ends]

2a.10. For each level of reliability testing 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.

[Response Begins]
Reliability/ Validity of the PROM-PHQ-9 and PHQ-9M
As PHQ-9 depression severity increased, there was a substantial decrease in functional status of all 6 SF-20 subscales in addition to an increase in symptom-related difficulty, sick days and health care utilization. Construct validity, using mental health professional re-interview as the criterion standard, has demonstrated a PHQ-9 score $\geq$ 10 has a sensitivity of 88% and a specificity of 88% for major depression. Additionally, a score <5 almost always signifies the absence of a depressive disorder, with a positive likelihood ratio of 0.04. Also, ROC analysis showed that the area under the curve for the PHQ-9 in diagnosing major depression was 0.95, suggesting a test that discriminates well between persons with and without major depression.
The internal reliability of the PHQ-9 was excellent, with Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study and 0.86 in the PHQ OBGYN Study. Test-retest reliability of the PHQ-9 was also excellent.
Correlation between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours was 0.84, and the mean scores were nearly identical (5.08 vs 5.03). [Validity of a Brief Depression Severity Measure Kronke, Kurt, Spitzer, Robert et al. J Gen Internal Medicine 2001 September; 16(9): 606–613. www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/]
In addition to the adults and elderly, the PHQ-9 has been validated in the adolescent populations (age 13 to 17). The PHQ-9M Modified for Teens is the PHQ-9 tool with slight word changes (in CAPS below) in three questions to modify the tool for the adolescent population with age appropriate terms.
Q2: Feeling down, depressed, IRRITABLE, or hopeless?
Q5: Poor appetite, WEIGHT LOSS, or overeating?
Q7: Trouble concentrating on things like SCHOOL WORK, reading, or watching TV?
Otherwise, the nine questions used in scoring the tool are identical to the PHQ-9. The copyright statement on the PHQ-9M tool states: *Modified with permission by the GLAD-PC team from the PHQ-9 (Spitzer, Williams & Kroenke, 1999), Revised PHQ-A (Johnson, 2002) and the CDS (DISC Development Group, 2000)*
Although widely used in pediatric practices and endorsed by the AAP, APA and AACAP, the modified version of the PHQ-9 tool has not had separate validation studies, as the nine questions are essentially the same as the original PHQ-9, which has been validated for adolescents ages 13 and older. The APA recommends using the modified version of the PHQ-9 for children ages 11 to 17 to assess depression symptom severity (APA, 2015). American Psychiatric Association. 2015. Online Assessment Measures. *Severity Measure for Depression, Child Age 11 to 17 (PHQ-9 modified for Adolescents [PHQ-A], Adapted)*. https://www.psychiatry.org/psychiatrists/practice/dsm/dsm-5/online-assessment-measures
Reliability of the PROM-PM:

Reliability is a function of provider-to-provider variation and samples size. Empirical testing of computed performance scores for reportable clinics was conducted using a beta-binomial model. Reliability ranges from 0.0 (no consistency) to 1.00 (perfect consistency). The extent to which the reliability falls below 1.00 is the extent to which errors of measurement are present. Reliability of 0.70 or greater is considered acceptable for drawing conclusions about groups.

- The BETABIN macro was used on each measure (SAS).
- Use the macro to get α and β.
- provider-to-provider variance: $\sigma^2 = (\alpha\beta) / (\alpha + \beta + 1)(\alpha + \beta)^2$
- plug this variance value into the reliability equation: $\sigma^2 / (\sigma^2 + (p(1 - p)/n))$
    - p = rate
    - n = number of eligible patients
- Determine reliability rate for each clinic.
- Average the reliability rate over all clinics.

2021 Submission

All results are stratified by adults and adolescents.

[Response Ends]

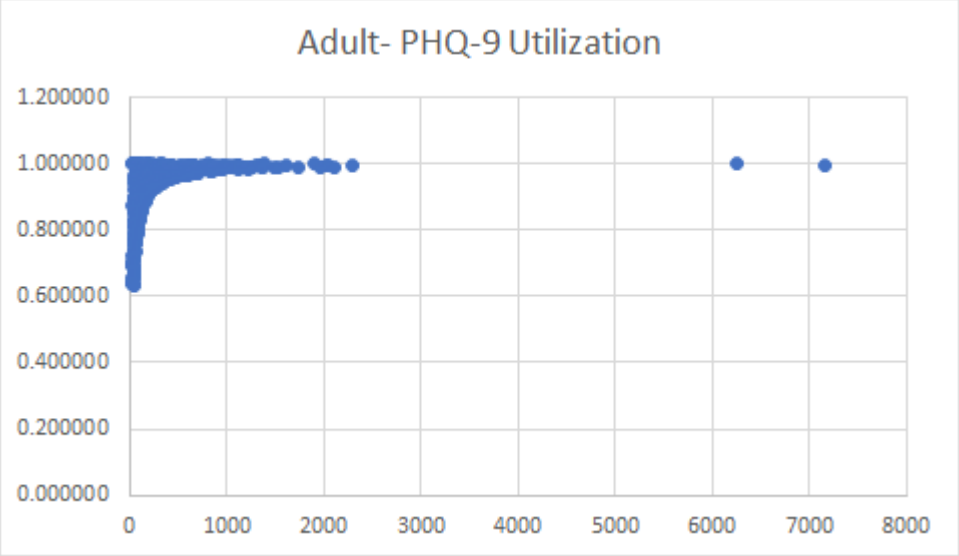
2a.11. For each level of reliability testing checked above, what were the statistical results from reliability testing?

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, NQF Measure Evaluation Criteria).

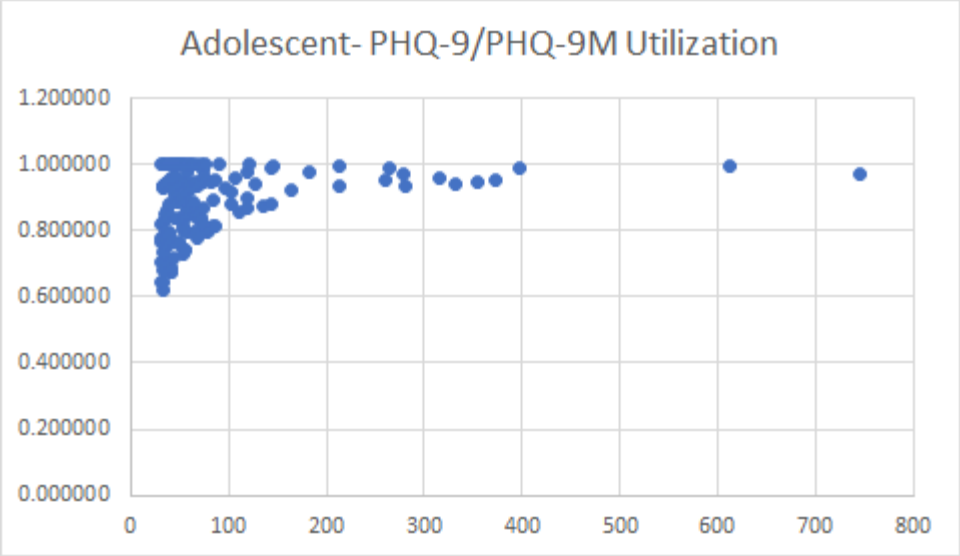
[Response Begins]

2021 Submission

PHQ-9 Assessment- Adults
601 clinics, 227,000 patients
Average Reliability Score: 0.932903



2020 Beta-binomial Reliability Performance Score- Adults 0.932903 (number of clinics 601, number of patients 227,127)
PHQ-9 Assessment- Adolescents
142 clinics, 12,616 patients
Average Reliability Score: 0.878959



2020 Beta-binomial Reliability Performance Score- Adolescents 0.878959 (number of clinics 142, number of patients 12,616

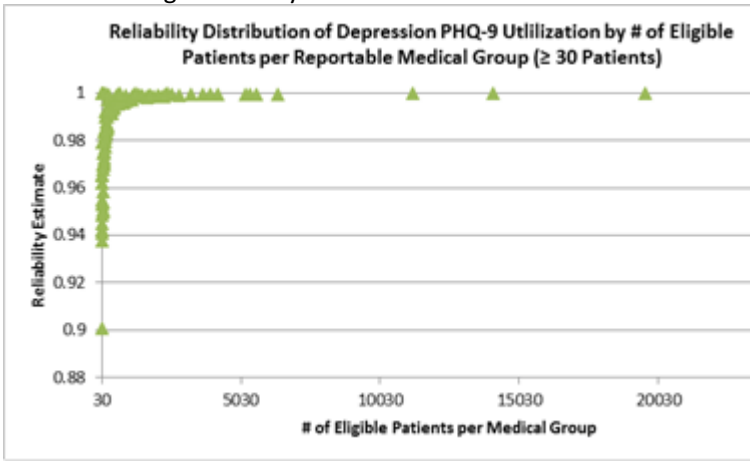
2013 Submission

Original Reliability Testing - Adults

Reliability = 0.987

Reportable medical groups (≥ 30 patients)

- $\alpha = 1.3292$
- $\beta = 0.9405$
- σ^2 (provider to provider variance) = 0.0742
- average reliability = 0.987



Adult patients age 18 and older; dates of service 10/1/2013 to 1/21/2014

[Response Ends]

2a.12. Interpret the results, in terms of how they demonstrate reliability.

(In other words, what do the results mean and what are the norms for the test conducted?)

[Response Begins]

PROM- PHQ-9

- PHQ-9 score > 10 has a sensitivity of 88% and a specificity of 88% for major depression.
- Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study and 0.86 in the PHQ OBGYN Study.
- PHQ-9M is only a slight modification of the original tool with developer's permission

The PHQ-9 patient reported outcome tool demonstrates sound psychometric properties (reliability, validity, specificity, and sensitivity to change) and is appropriate for measuring patient outcomes related to depression.

The PRO-PM Measure:
Clinic level reliability statistics are stratified by adult patients age 18 and older and adolescent patients age 12 to 17.
2021 Submission

- Reliability score = 0.932903 (Adult) and 0.878959 (Adolescents)

For clinics reporting measure results for adults (601 clinics and 227,127 patients), the reliability performance score was calculated at 0.932903. A beta-binomial reliability (signal-to-noise) score of greater than 0.70 indicates that it is acceptable to draw conclusions about groups, in this case by the comparison of clinic site level reporting. With a reliability score exceeding 0.93, there is the ability to distinguish higher performing clinics from lower performing clinics.
Although there are fewer clinics reporting measure results for adolescents (142) and fewer adolescents (12,616) as compared to the adult population, the reliability performance score is still quite high at 0.878959. This demonstrates that for the adolescent population, results can be used to distinguish higher performing clinics from lower performing clinics.
This data analysis, along with precise specifications and excellent validation results of critical data elements, demonstrates this measure construct to be reliable and detects meaningful differences among provider groups.

[Response Ends]

2b.01. Select the level of validity testing that was conducted.

[Response Begins]
Patient or Encounter-Level (data element validity must address ALL critical data elements)
Accountable Entity Level (e.g. hospitals, clinicians)
Empirical validity testing
[Response Ends]

2b.02. 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.

[Response Begins]
Reliability/ Validity of the PROM- PHQ-9:
As PHQ-9 depression severity increased, there was a substantial decrease in functional status of all 6 SF-20 subscales in addition to an increase in symptom-related difficulty, sick days and health care utilization. Construct validity, using mental health professional re-interview as the criterion standard, has demonstrated a PHQ-9 score > 10 has a sensitivity of 88% and a specificity of 88% for major depression. Additionally, a score <5 almost always signifies the absence of a depressive disorder, with a positive likelihood ratio of 0.04. Also, ROC analysis showed that the area under the curve for the PHQ-9 in diagnosing major depression was 0.95, suggesting a test that discriminates well between persons with and without major depression.
The internal reliability of the PHQ-9 was excellent, with Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study and 0.86 in the PHQ OBGYN Study. Test-retest reliability of the PHQ-9 was also excellent. Correlation between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours was 0.84, and the mean scores were nearly identical (5.08 vs 5.03).
[Validity of a Brief Depression Severity Measure Kronke, Kurt, Spitzer, Robert et al. J Gen Internal Medicine 2001 September; 16(9): 606–613. www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/
Validity of the PROM-PM:
Data Element Validity: Validating the submitted data via the direct data submission process is completed in four steps: denominator certification, data quality checks, validation audit, and the two-week medical group review period.
Pre-submission certification occurs prior to data collection and extraction/abstraction ensures that all medical groups apply the denominator criteria correctly and in a consistent manner. MNCM staff review the documentation to verify all criteria were applied correctly, prior to approval for data submission.
Denominator certification documentation for this measure includes:

- Date of Birth (ranges)
- Date of Service (ranges)
- ICD-10 Codes used
- Exclusions to the measure and attest to mechanism to submit exclusion code/ reason for exclusion reasons that may happen after a patient has an index contact.

Groups additionally supply their query code for review.
Following data submission to the MNCM Data Portal there are additional data quality checks in place for evaluating the accuracy of data submitted. During file upload, program checks for valid dates, codes and values and presents users with errors and warnings. Additionally, MNCM staff review population counts (denominator) and outcome rates for any significant variance from the previous year's submission and may prompt further clarification from the medical group.
Validity Performance Score: Correlation was performed against a depression outcome measure; the hypothesis tested was that clinics that do well assessing their patients with a diagnosis of depression frequently with the PHQ-9/ PHQ-9M will also perform better in achieving remission (PHQ-9 < 5) at six months.

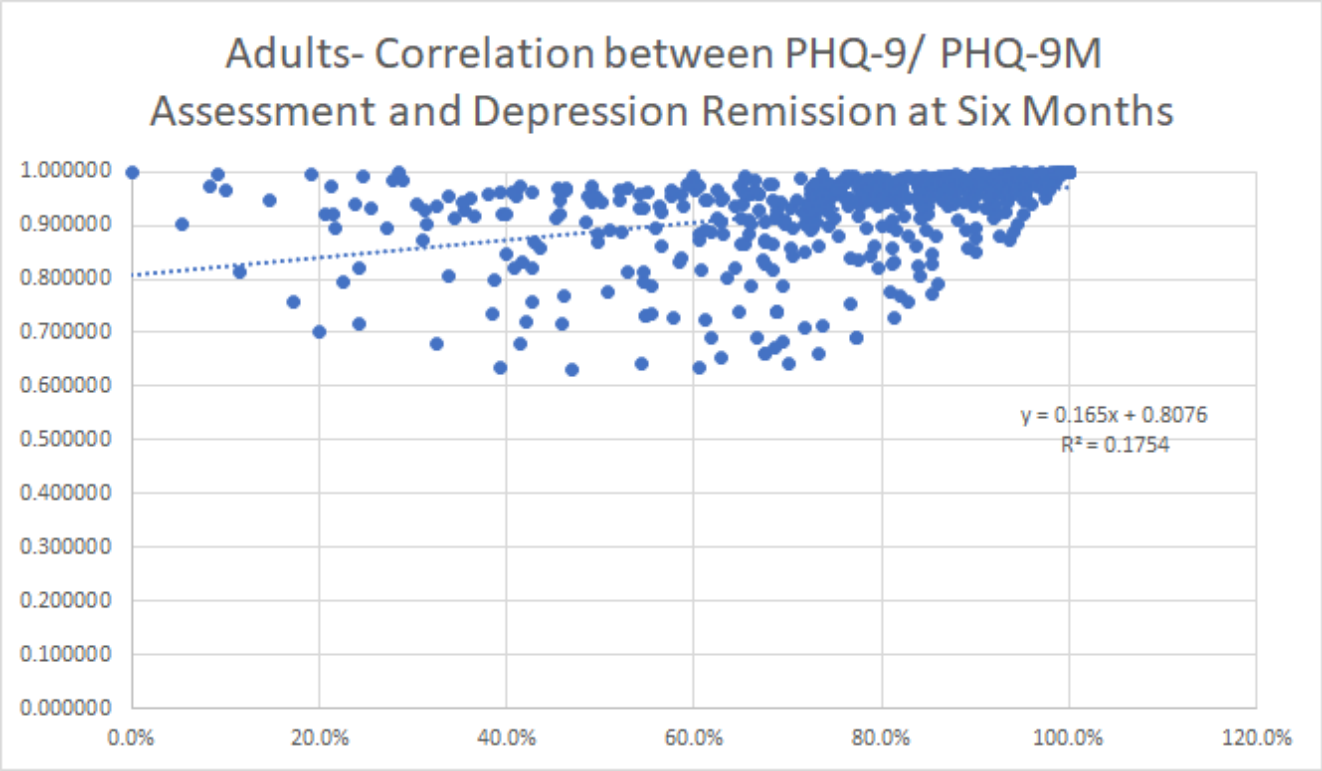
1. Perfect: If the value is near ± 1, then it is said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2. High degree: If the coefficient value lies between ± 0.50 and ± 1, then there is said to be a strong correlation.

[Response Ends]

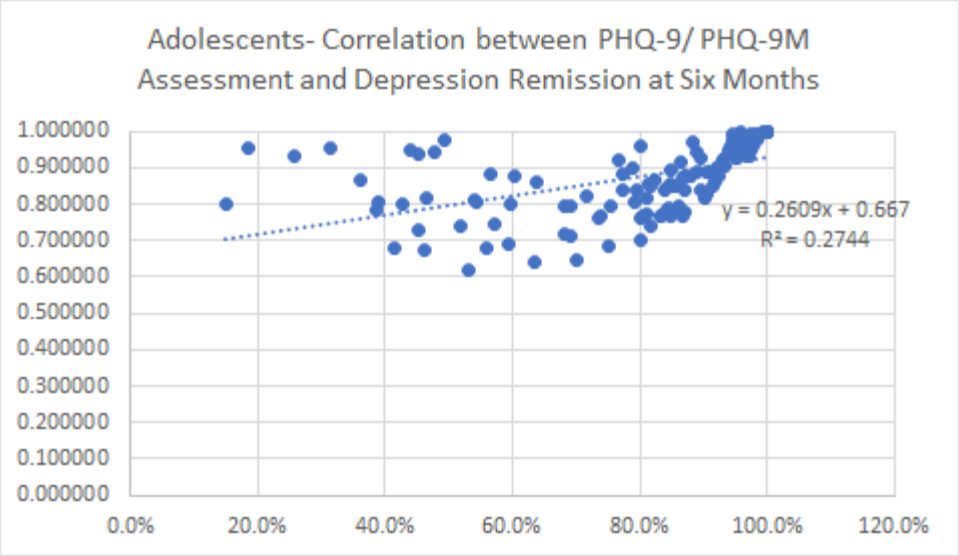
2b.03. Provide the statistical results from validity testing.

Examples may include correlations or t-test results.

[Response Begins]
2021 Submission
Validity Performance Score (Correlation)- Adults



Adults correlation between assessment with PHQ-9/PHQ-9M and Depression Remission at Six months; R-squared 0.1754
Validity Performance Score (Correlation)- Adolescents



Adolescents correlation between assessment with PHQ-9/PHQ-9M and Depression Remission at Six months; R-squared 0.2744

2020 Validation Summary- Data Elements

Pre-Submission	Post-submission Data Quality Checks	Audit of Data Source
49% of groups passed with no errors. Types of errors: dates of service, dates of birth, ICD-10 codes, exclusions not applied correctly, intended to submit only one screening per patient Typically, most groups are able to correct file extraction issues, but this year eight groups did not proceed with correction and submission, citing EMR changes, resource limitations and inabilities related to prioritization during the COVID-19 pandemic.	58% of those that submitted data passed initial quality checks. Types of errors: insurance data, RELC data, file formatting that caused improper rate calculation (dx codes with extra spaces or no decimals), transposed counts for adult and adolescent populations, inability to submit full dates of service for the adolescent population, inconsistent patient ID format which impacted indexing and outcomes, incorrect dates of service/dates of birth Three groups did not proceed with correction of their submission, citing EMR changes, resource limitations and inabilities related to prioritization during the COVID-19 pandemic.	**30% of groups that submitted data were audited; 94% passed the audit.** Types of errors: file formatting produced incorrect PHQ-9 scores, inconsistent patient IDs

[Response Ends]

2b.04. Provide 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?)

[Response Begins]

The PHQ-9/PHQ-9M patient reported outcome tool demonstrates sound psychometric properties (reliability, validity, specificity and sensitivity to change) and is appropriate for use in the assessment of patients with depression. There was high compliance with critical data element validity as demonstrated by annual validation audit processes.

The adult and adolescent stratifications demonstrate a fairly weak correlation in a positive direction [R squared 0.1754 and 0.2744 respectively] against a related outcome measure. Despite the hypothesis that clinics whose patients with depression are regularly assessed would achieve higher rates of remission at six months, this proved to not be a strong correlation. However, it is important to continually assess patients with a current diagnosis of depression or a history of depression for depression symptoms. The measure is related to and supports the outcome measures of depression remission and response at six and twelve months. (NQF # 0710, 0711, 1884 and 1885)

[Response Ends]

2b.05. 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 in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]

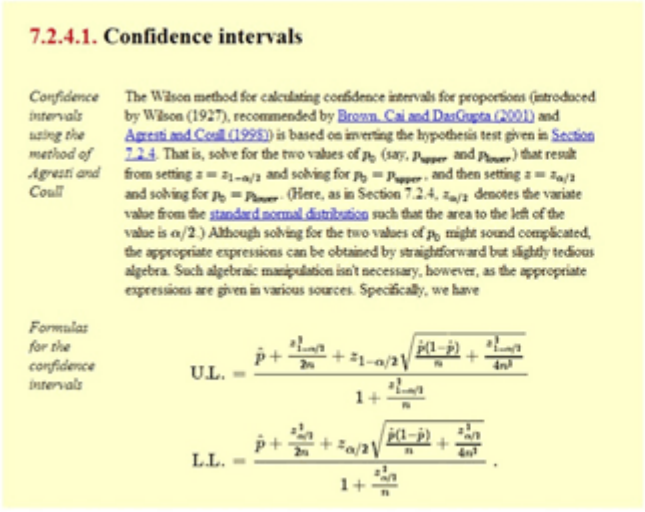
DATA COLLECTION

Data are reported at two levels: by clinic site and medical group. Clinic abstractors collect data from medical records either by extracting the data from an electronic medical record (EMR) via data query or from abstraction of paper-based medical records. All appropriate Health Insurance Portability and Accountability (HIPAA) requirements are followed for data transfer to MNCM.

MNCM staff conduct an extensive validation process including pre-submission data certification, post submission data quality checks of all files, and audits of the data source for selected clinics. For medical record audits, MNCM uses NCQA's "8 and 30" File Sampling Procedure, developed in 1996 in consultation with Johns Hopkins University. For a detailed description of this procedure, see www.ncqa.org. Audits are conducted by trained MNCM auditors who are independent of medical groups and/or clinics. The validation process ensures the data are reliable, complete and consistent.

ELIGIBLE POPULATION SPECIFICATIONS The eligible population for each measure is identified by a medical group on behalf of their individual clinics. MNCM's 2019 DDS Data Collection Guides provide technical specifications for the standard definitions of the eligible population, including elements such as age.

NUMERATOR SPECIFICATIONS For DDS measures, the numerator is the number of patients identified from the eligible population who meet the numerator criteria. The numerator is calculated using the clinical quality data submitted by the medical group; this data is verified through MNCM's validation process



7.2.4.1. Confidence intervals

Confidence intervals using the method of Agresti and Coull

The Wilson method for calculating confidence intervals for proportions (introduced by Wilson (1927), recommended by Brown, Cai and DasGupta (2001) and Agresti and Coull (1998)) is based on inverting the hypothesis test given in Section 7.2.4. That is, solve for the two values of p_0 (say, p_{upper} and p_{lower}) that result from setting $z = z_{1-\alpha/2}$ and solving for $p_0 = p_{upper}$, and then setting $z = z_{\alpha/2}$ and solving for $p_0 = p_{lower}$. (Here, as in Section 7.2.4, $z_{\alpha/2}$ denotes the variate value from the standard normal distribution such that the area to the left of the value is $\alpha/2$.) Although solving for the two values of p_0 might sound complicated, the appropriate expressions can be obtained by straightforward but slightly tedious algebra. Such algebraic manipulation isn't necessary, however, as the appropriate expressions are given in various sources. Specifically, we have

Formulas for the confidence intervals

$$\text{U.L.} = \frac{\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{z^2_{1-\alpha/2}}{n}}$$

$$\text{L.L.} = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + \frac{z^2_{\alpha/2}}{n}}.$$

The Wilson method for calculating confidence intervals for all clinic rates and statewide rates.

www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

Equation for the Calculation of Confidence Intervals; Wilson Method

CALCULATING RATES

Due to the dynamic nature of patient populations, rates and 95 percent confidence intervals are calculated for each measure for each medical group/clinic regardless of whether the full population or a sample is submitted. The statewide average rate is displayed when comparing a single medical group/clinic to the performance of all medical groups/clinics to provide context. The statewide average is calculated using all data submitted to MNCM which may include some data from clinics located in neighboring states.

THRESHOLD FOR PUBLIC REPORTING

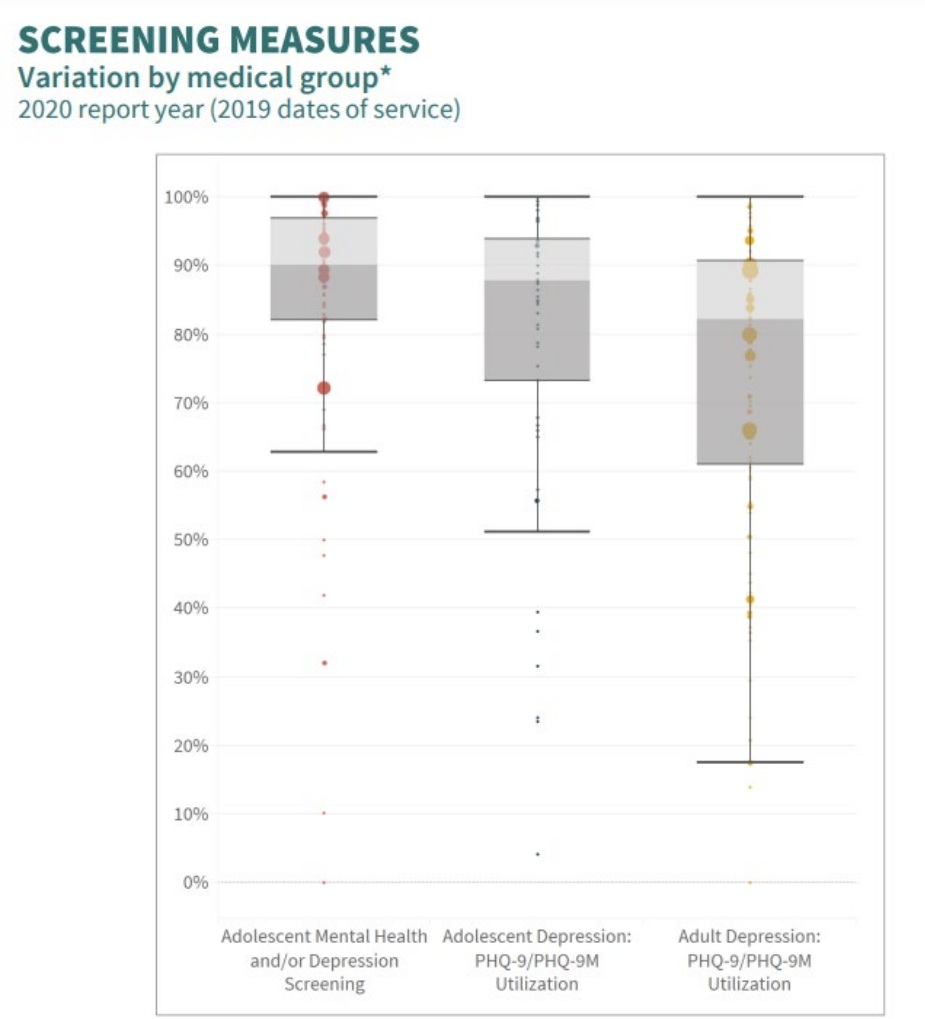
MNCM has established minimum thresholds for public reporting of DDS measures to ensure statistically reliable rates. Only medical groups and clinics that meet the threshold of 30 patients in the denominator of each measure are publicly reported.

[Response Ends]

2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.

Examples may include 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.

[Response Begins]
DEPRESSION CARE IN MINNESOTA: ADULTS & ADOLESCENTS 2020 REPORT YEAR (2019 DATES OF SERVICE)
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Report.pdf
Variability is demonstrated by box plot quartiles demonstrating outliers, the minimum and maximum values, upper quartile, median and lower quartile. Distribution of rates demonstrates variability and opportunity for improvement.



Screening Measures; Depression Care in Minnesota 2020 Report Year. This image depicts the variability of rates among medical groups around the statewide average: 77.7% for Adults (n = 248,163) 78.4% for Adolescents (n = 19,574). Variability among medical groups is displayed by the range of results (25 to 100% and 8 to 100% respectively. Box plot diagrams further display the wide variability in medical group rates; for the adults, a significant portion of clinics are in the lower quartile.
The Adult Depression: PHQ-9/ PHQ-9M Utilization measure has the widest variation among medical groups. The Adolescent Mental Health and/or Depression Screening measure has the most consistent performance among medical groups.

The image above depicts the variability of rates among medical groups around the statewide average:

- 77.7% for Adults (n = 248,163)
- 78.4% for Adolescents (n = 19,574).

Variability among medical groups is displayed by the range of results (25 to 100% and 8 to 100% respectively. Box plot diagrams further display the wide variability in medical group rates; for the adults, a significant portion of clinics are in the lower quartile.

Adult Depression- PHQ-9/PHQ-9M Utilization
Partial Alphabetical Listing of Minnesota Medical Groups with Ranking of Results
Full list in report at
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
STATEWIDE AVERAGE	*	248,162	77.7%	*	*
AALFA Family Clinic	Above Average	137	95.6%	90.8%	98.0%
Advanced Medical Clinic q	Below Average	102	20.6%	13.9%	29.4%
Allina Health	Above Average	28,911	89.2%	88.8%	89.5%
Allina Health Apple Valley	Above Average	648	87.8%	85.1%	90.1%
Alomere Health	Above Average	1,191	95.0%	93.6%	96.1%
Altru Health System	Below Average	3,079	17.5%	16.2%	18.9%
Amery Hospital and Clinic	Above Average	635	92.1%	89.8%	94.0%
Appleton Area Health Services	Above Average	156	88.5%	82.5%	92.6%
Associated Clinic of Psychology	Above Average	2,694	85.7%	84.3%	87.0%
Bluestone Physician Services	Below Average	1,106	24.0%	21.5%	26.6%
Boynton Health Service	Above Average	2,216	98.5%	97.9%	98.9%
Catholic Charities Behavioral Health Clinic	Top Performer	172	100.0%	97.8%	100.0%
CCM Health	Top Performer	855	100.0%	99.6%	100.0%
Cedar Riverside People's Center	Above Average	162	84.6%	78.2%	89.3%
CentraCare Health	Above Average	10,759	93.7%	93.2%	94.1%
Children's Minnesota	Top Performer	217	100.0%	98.3%	100.0%
Cromwell Medical Clinic PLLC - IHN	Below Average	83	55.4%	44.7%	65.6%
Cuyuna Regional Medical Center	Above Average	681	93.2%	91.1%	94.9%
Dawson Clinic	Below Average	405	43.7%	39.0%	48.6%
Dr. VM Baich, PA	Average	32	81.3%	64.7%	91.1%
Duluth Family Medicine Clinic	Below Average	358	61.5%	56.3%	66.3%
Edina Sports Health & Wellness	Below Average	529	42.2%	38.0%	46.4%
Entira Family Clinics	Average	3,578	78.9%	77.5%	80.2%
Essentia Health	Below Average	15,090	65.5%	64.7%	66.3%
Fairview Health Services	Above Average	22,452	79.9%	79.4%	80.4%
Fairview Mesaba Clinics	Above Average	1,243	83.3%	81.1%	85.2%
Family Practice Medical Center of Willmar	Above Average	214	97.2%	94.0%	98.7%
France Avenue Family Physicians	Above Average	867	97.0%	95.6%	97.9%
Glencoe Regional Health Services	Above Average	616	95.8%	93.9%	97.1%
Glenwood Medical Center	Below Average	948	13.8%	11.8%	16.2%
Grand Itasca Clinic	Above Average	742	83.2%	80.3%	85.7%
Granite Falls Municipal Hospital	Below Average	287	20.9%	16.6%	26.0%
Gundersen Health System	Above Average	412	84.7%	80.9%	87.9%

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
Gundersen Saint Elizabeth's	Top Performer	169	100.0%	97.8%	100.0%
HealthPartners Central Minnesota	Above Average	862	89.4%	87.2%	91.3%
Health Partners Clinics	Above Average	12,718	89.9%	89.4%	90.4%

* Cell intentionally left empty

Ranking of Clinics Alphabetical with Health Score Ranking
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf page 5

The image above is a portion of a report the demonstrates the public reporting and display the Health Score Rankings for each medical group for this measure for adults.

Adolescent Depression- PHQ-9/PHQ-9M Utilization
Partial Alphabetical Listing of Minnesota Medical Groups with Ranking of Results
Full list in report at
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
STATEWIDE AVERAGE	*	19,574	78.4%	*	*
Allina Health	Above Average	1,751	92.9%	91.6%	94.0%
Alomere Health	Above Average	86	98.8%	93.7%	99.8%
Altru Health System	Below Average	321	31.5%	26.6%	36.7%
Amery Hospital and Clinic	Above Average	47	93.6%	82.8%	97.8%
Associated Clinic of Psychology	Average	131	84.7%	77.6%	89.9%
CentraCare Health	Above Average	1,171	96.4%	95.2%	97.3%
Children's Minnesota	Top Performer	532	100.0%	99.3%	100.0%
Cuyuna Regional Medical Center	Above Average	43	93.0%	81.4%	97.6%
Eagan Valley Pediatrics	Top Performer	53	100.0%	93.2%	100.0%
Entira Family Clinics	Above Average	139	89.9%	83.8%	93.9%
Essentia Health	Above Average	1,290	84.3%	82.3%	86.2%
Fairview Health Services	Above Average	848	87.6%	85.2%	89.7%
Fairview Mesaba Clinics	Above Average	63	88.9%	78.8%	94.5%
Grand Itasca Clinic	Above Average	48	93.8%	83.2%	97.9%
HealthPartners Central Minnesota	Above Average	61	96.7%	88.8%	99.1%
Health Partners Clinics	Above Average	593	91.9%	89.4%	93.8%
Hennepin Healthcare	Above Average	715	99.4%	98.6%	99.8%
Hutchinson Health	Below Average	87	67.8%	57.4%	76.7%
Lake Region Healthcare	Above Average	131	91.6%	85.6%	95.2%
Lakeland Mental Health	Below Average	295	57.3%	51.6%	62.8%
Lakewood Health System	Average	118	81.4%	73.4%	87.4%
Mankato Clinic	Above Average	342	96.5%	94.0%	98.0%
Mayo Clinic	Average	869	78.7%	75.9%	81.3%
Mayo Clinic Health System	Above Average	784	86.5%	83.9%	88.7%

* Cell intentionally left empty

Ranking of Clinics Alphabetical with Health Score Ranking
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf page 9

[Response Ends]

2b.07. Provide 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.

In other words, what do the results mean in terms of statistical and meaningful differences?

[Response Begins]
With statewide averages of 78% for both populations, this measure continues to demonstrate opportunity for improvement in the frequent assessment of depression symptoms and identifies meaningful differences among providers.

*	Quality Measure	2020 RY Statewide Average (2019 Dates of Service)	Total Eligible Patients (Denominator)
Screening Measures	Adolescent Mental Health and/or Depression Screening	88.7%	166,311
Screening Measures	Adolescent Depression: PHQ-9/PHQ-9M Utilization	78.4%	19,574
Screening Measures	Adult Depression: PHQ-9/PHQ-9M Utilization	77.7%	248,162
Six Month Measures	ADULTS	*	*
Six Month Measures	Adult Depression: Six Month Follow-up	48.5%	120,344
Six Month Measures	Adult Depression: Response at Six Months	19.4%	120,344
Six Month Measures	Adult Depression: Remission at Six Months	11.3%	120,344
Six Month Measures	ADOLESCENTS	*	*
Six Month Measures	Adolescent Depression: Six Month Follow-up	43.4%	11,658
Six Month Measures	Adolescent Depression: Response at Six Months	15.5%	11,658
Six Month Measures	Adolescent Depression: Remission at Six Months	8.0%	11,658
12 Month Measures	ADULTS	*	*
12 Month Measures	Adult Depression: 12 Month Follow-up	41.8%	120,344
12 Month Measures	Adult Depression: Response at 12 Months	17.0%	120,344
12 Month Measures	Adult Depression: Remission at 12 Months	10.1%	120,344
12 Month Measures	ADOLESCENTS	*	*
12 Month Measures	Adolescent Depression: 12 Month Follow-up	38.9%	11,658
12 Month Measures	Adolescent Depression: Response at 12 Months	14.5%	11,658
12 Month Measures	Adolescent Depression: Remission at 12 Months	7.8%	11,658

* Cell intentionally left empty

This table provides an overview of the statewide rates for the mental health measures.
While screening for mental health/depression continue to increase, outcomes among patients with depression (i.e., response and remission) continues to show room for improvement for both adults and adolescents.
Statewide average:
The average performance rate among medical groups for the 2020 report year.

Summary of Depression Measures; Rates and Denominator
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Report.pdf
The image above is a portion of a report for illustrative purposes to display the public reporting of rates and denominator counts for this measure. Of note: Adults with a rate of 77.7% for 248,162 patients and Adolescents with a rate of 78.4% for 19,574 patients.

[Response Ends]

2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]

[Response Begins]
Though it is well recognized that maintaining ongoing contact with this population of patients with depression is critical to their successful remission of symptoms, it is also very challenging to do so. Of any patient population, patients with depression are least likely to be able to self-advocate and require processes and systems in place for maintaining contact. MN has made improvements in rates of assessment, from 55.4% in 2010 to 77.7% in 2020 for adults.
Missing data is not an issue for this measure as patients with a diagnosis of major depression or dysthymia who have a visit or contact within the measurement period who are not assessed at least once in the four month period remain in the denominator.
This measure is a companion related measure that allows medical groups to understand their use of the PHQ-9/ PHQ-9M tool in assessing depression and related to remission and response outcome measures (NQF #s 0710, 0711, 1884 and 1885)

[Response Ends]

2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).

[Response Begins]
Missing data is not an issue. Patients who are not assessed with a PHQ-9/ PHQ-9M at least once during a visit or contact within a four month measurement period are included in the denominator.
[Response Ends]

2b.10. Provide 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.

In other words, 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 was conducted, justify the selected approach for missing data.

[Response Begins]
Missing data is not an issue for this measure as constructed; please see discussion in 2b.08
[Response Ends]

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 eCQMs). 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.

2b.11. Indicate whether there is more than one set of specifications for this measure.

[Response Begins]
No, there is only one set of specifications for this measure
[Response Ends]

2b.12. 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. Indicate what statistical analysis was used.

[Response Begins]
[Response Ends]

2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.

Examples may include correlation, and/or rank order.

[Response Begins]
[Response Ends]

2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.

In other words, what do the results mean and what are the norms for the test conducted.

[Response Begins]
[Response Ends]

2b.15. Indicate whether the measure uses exclusions.

[Response Begins]
Yes, the measure uses exclusions.
[Response Ends]

2b.16. Describe the method of testing exclusions and what was tested.

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?

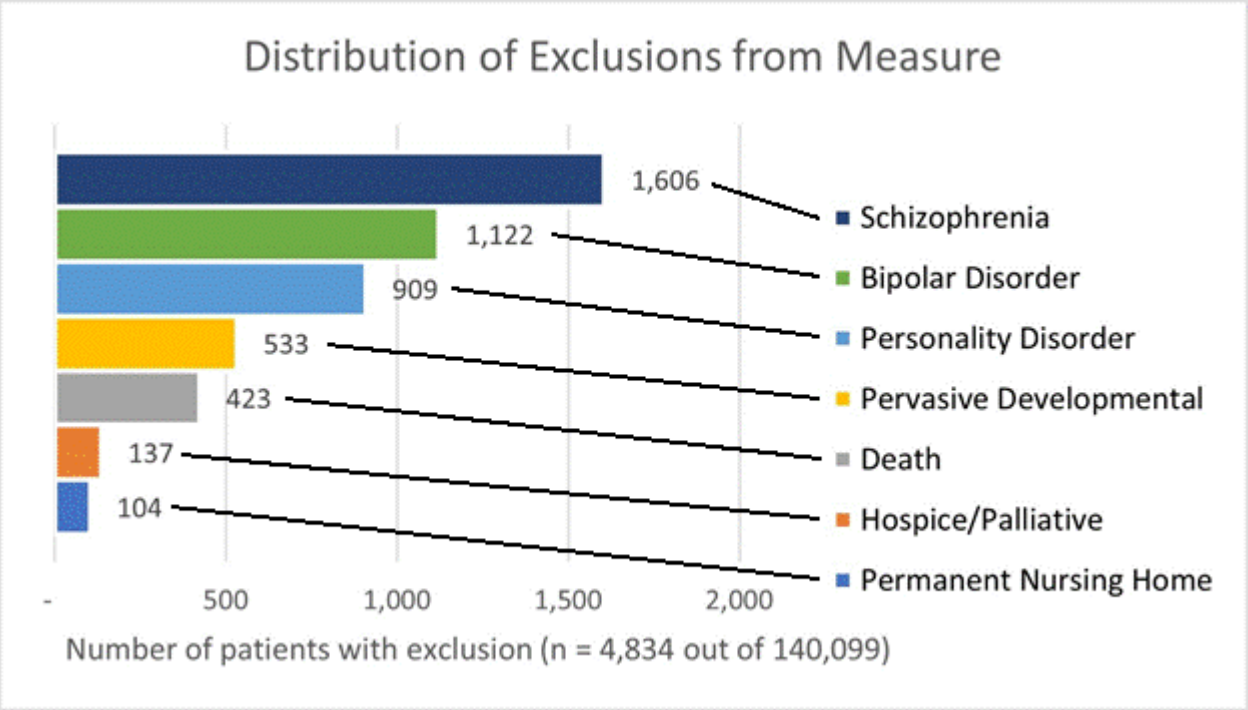
[Response Begins]
Exclusions for this process measure paired with outcome measures of depression remission and response are harmonized (match exclusions for the outcome measures). Rationale for exclusions are of a clinical nature where expectations for outcomes may be different due to life expectancy (nursing home residents, hospice/ palliative care, death) or co-morbid diagnoses that may emerge after initial impression/ diagnosis of a depressive disorder (bipolar or personality disorder). Also need a mechanism to exclude bipolar disorder patients who frequently also have diagnosis of major depression despite this being a departure from best coding practices.

[Response Ends]

2b.17. Provide 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.

[Response Begins]
2013- When known, exclusions are removed "up-front", prior to data submission and validated through the denominator certification process and these exclusions are not available for analysis. When exclusions occur after the index contact event, they are included in the data submission for this measure and are available for analysis. 97.0% of the eligible patients remain in the denominator without need for further exclusion because of events or diagnoses occurring after index. Of the 3% of the population that do require exclusion after index, 86% were because of diagnosis of bipolar or personality disorder and 14% due to death, hospice or permanent nursing home residence.



Distribution of Exclusions of Patients with a Diagnosis of Major Depression or Dysthymia. Rate of 3.45%
The above image is a stacked bar chart demonstrating the frequency of exclusions used for a population of 140,099 patients. The most frequently occurring exclusion is schizophrenia (blue bar) followed by bipolar disorder (green bar). This is not a surprising result because clinically, these two conditions can have a depressive component. However, their treatments and outcomes are very different from major depression, and they represent appropriate exclusions from the measure.

[Response Ends]

2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.

In other words, 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.

[Response Begins]
Depression, like many chronic or episodic conditions, does not often exist in isolation from other medical conditions. Some mental health conditions like bipolar disorder or schizophrenia can have a component of depression or occur concurrently, but patients with these conditions have very different outcomes and to include them would distort the result of the measure. The goals related to measure development in terms of exclusions are to be patient centered and as inclusive as possible without distortion of the measure results.
Overall, exclusions do not limit or reduce the desired target population of patients with major depression or dysthymia. Updated analysis of modifications and additions to exclusions demonstrate continued appropriate clinical indication without reducing the target population.

[Response Ends]

2b.19. Check all methods used to address risk factors.

[Response Begins]
No risk adjustment or stratification
[Response Ends]

2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.

[Response Begins]
Not applicable
[Response Ends]

2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.

[Response Begins]
Not applicable
[Response Ends]

2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]
Other (specify)
[Other (specify) Please Explain]
Not applicable, process measure is not risk-adjusted

[Response Ends]

2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10 or other statistical tests; correlation of x or higher. Patient factors should be present at the start of care, if applicable. Also discuss any "ordering" of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).

[Response Begins]
Not applicable
[Response Ends]

2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.

[Response Begins]
Not applicable
[Response Ends]

2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.

Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.

[Response Begins]
Not applicable

[Response Ends]

2b.26. 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 control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter "N/A" for questions about the statistical risk model discrimination and calibration statistics.

Validation testing should be conducted in a data set that is separate from the one used to develop the model.

[Response Begins]
Not applicable
[Response Ends]

2b.27. Provide risk model discrimination statistics.

For example, provide c-statistics or R-squared values.

[Response Begins]
Not applicable
[Response Ends]

2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).

[Response Begins]
Not applicable
[Response Ends]

2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.

The preferred file format is .png, but most image formats are acceptable.

[Response Begins]
Not applicable
[Response Ends]

2b.30. Provide the results of the risk stratification analysis.

[Response Begins]
Not applicable
[Response Ends]

2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).

In other words, what do the results mean and what are the norms for the test conducted?

[Response Begins]
Not applicable
[Response Ends]

2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.

Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.

[Response Begins]
Not applicable
[Response Ends]

Criteria 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.

3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.

[Response Begins]
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score)
Coded by someone other than person obtaining original information (e.g., DRG, ICD-10 codes on claims)
[Response Ends]

3.02. Detail to what extent the specified data elements are available electronically in defined fields.

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.
[Response Begins]
ALL data elements are in defined fields in electronic health records (EHRs)
[Response Ends]

3.03. 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 data elements not from electronic sources.

[Response Begins]
not applicable
[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]
This measure was developed as an e-CQM (legacy measure) and one of the first adopted into CMS' Measure Authoring Tool (MAT), CMS 160 8.4 and was used for several years in the e-CQM program until it was recommended for removal from the MIPS program by CMS as part of the 2020 rule making process (effective MIPS Payment Year 2022). Rationale indicated favor for the more robust companion outcome measure Depression Remission at Twelve Months (Q370/CMS159/NQF 0710e)
[Response Ends]

3.06. 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.

[Response Begins]
MNCM has developed a direct data submission process in 2006, whereby medical groups submit a patient level data file of a minimal data set (only those elements needed for measure calculation, risk adjustment and stratification/ analysis) to our HIPAA secure data portal for rate calculation and public reporting. Utilizing the direct data submission process we have learned the following:

1. Data Submission- Providing data collection software for medical groups wishing to submit data was not always the best and most efficient way of collecting data. As electronic health records use becomes more pervasive in our state, providing templates of data file submissions proved to be more efficient.
2. Specifications- Detailed specifications with instructions on how to handle most situations (e.g. detailed instructions on blood pressure values) has been valuable to medical groups, increased data accuracy is reflected by 98-99% of medical groups meeting validation standards for submitted data against the medical record.

3. Audit- Audit methods have insured the accuracy of our data and we are able to successfully compare providers because everyone is pulling their data the same way and subject to the same rules.
4. Confidentiality- Patient confidentiality has been addressed by numerous mechanisms. MNCM only receives the patient level information needed to calculate the rates, determine eligibility for inclusion in the measure and support the administration of pay for performance programs. The PHI submitted is minimal and the data is protected by 1) password protection with password only available to the medical group submitting data, 2) file upload process is encrypted as data is transferred and 3) Data is stored on a separate secure server and meets all HIPAA protection rules.
5. Acceptance of Data- Vast improvement in terms of the timeliness of the data submitted by medical groups six weeks after the end of the measurement period as compared to prior method of health plan's samples and the results over a year old. Providers are more accepting of the results as compared to previous methods of pooling health plan samples.
6. Data Collection Burden- We have learned that for additional future measures we will need to stagger the data collection time frames and submission deadlines as to not burden the medical groups in terms of abstraction/ extraction.
7. Health Plans: pay for performance and the inclusion of measures within contracts significantly impacts the number of groups participating in each measure.
8. Patient Reported Outcome (PROM) assessment tools. Consideration for inclusion of a PROM includes the following: a tool that is psychometrically sound (valid/ reliable/ specific and sensitive to change), providers are amenable to the use of the tool, can be implemented into clinical work flows, can be administered by multiple modes including electronic administration and tool is valuable to patients and does not cause undue completion burden.

MNCM is implementing a new data collection method, PIPE (Process Intelligence Performance Engine) that serves as a warehouse of clinical data (encounters, problem lists, labs, medications, etc) where measures are calculated centrally, significantly reducing data collection burden for providers.
https://helpdesk.mncm.org/helpdesk/KB/View/32539666-a-new-approach-to-measurement-introduction-to-pipe-recorded-presentation-and-slide-deck

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

3.07. Detail 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),

Attach the fee schedule here, if applicable.

[Response Begins]
No fees associated with the PROMs; PHQ-9 is publicly available at www.phqscreeners.com and PHQ-9M at https://www.aacap.org/App_Themes/AACAP/docs/member_resources/toolbox_for_clinical_practice_and_outcomes/symptoms/GLAD-PC_PHQ-9.pdf. In MN, no fees for data submission and rate calculation, however groups do incur the costs of data collection/ extraction/ abstraction needed to submit data.
No fees associated with the PIPE system.

[Response Ends]

Criteria 4: Use and Usability

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.

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.
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 demonstrating performance improvement.

4a.01. Check all current uses. For each current use checked, please provide:

Name of program and sponsor

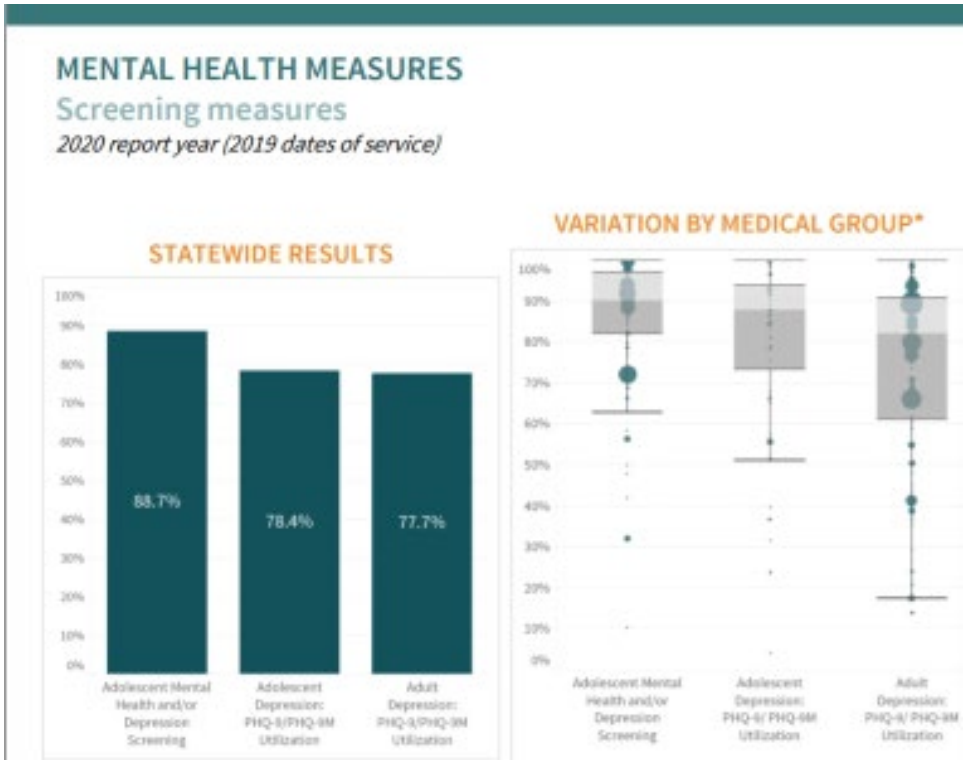
URL

Purpose

Geographic area and number and percentage of accountable entities and patients included

Level of measurement and setting

[Response Begins]
Public Reporting
[Public Reporting Please Explain]
MN Community Measurement- a non-profit 501 (c)(3) whose mission is to accelerate the improvement of health by publicly reporting health care information. Founding members include: Blue Cross and Blue Shield of MN, HealthPartners, Medica, Metropolitan Health Plan, Minnesota Medical Association, Minnesota Hospital Association, PreferredOne, PrimeWest Health System, South Country Alliance and Ucare Minnesota.
Geographic Area:
All primary care clinics in MN and bordering communities in Wisconsin, North Dakota, South Dakota and Iowa.
MN HealthScores A consumer facing public reporting website at www.mnhealthscores.org
Rates for this measure are published annually in the MNCM Health Care Quality report at
https://mncm.org/reports/#community-reports.



Example of Public Reporting; Annual Health Care Quality Report

Rates for this measure are also published annually on the consumer facing public website MN HealthScores https://www.mnhealthscores.org/ however, due to redesign of this depression measure to incorporate adolescents and a currently in progress website redesign, it is not available on MN HealthScores.

Quality Improvement with Benchmarking (external benchmarking to multiple organizations)

[Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Please Explain]

Rates for this measure are published annually in the MNCM Health Care Quality report at https://mncm.org/reports/#community-reports. Appendices for this report display the measure by all medical groups in Minnesota alphabetically with an indication if performance is above, at or below the statewide average, snippet included:

Adult Depression- PHQ-9/PHQ-9M Utilization

Partial Alphabetical Listing of Minnesota Medical Groups with Ranking of Results

Full list in report at

https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
STATEWIDE AVERAGE	*	248,162	77.7%	*	*
AALFA Family Clinic	Above Average	137	95.6%	90.8%	98.0%
Advanced Medical Clinic q	Below Average	102	20.6%	13.9%	29.4%
Allina Health	Above Average	28,911	89.2%	88.8%	89.5%
Allina Health Apple Valley	Above Average	648	87.8%	85.1%	90.1%
Alomere Health	Above Average	1,191	95.0%	93.6%	96.1%
Altru Health System	Below Average	3,079	17.5%	16.2%	18.9%
Amery Hospital and Clinic	Above Average	635	92.1%	89.8%	94.0%
Appleton Area Health Services	Above Average	156	88.5%	82.5%	92.6%
Associated Clinic of Psychology	Above Average	2,694	85.7%	84.3%	87.0%
Bluestone Physician Services	Below Average	1,106	24.0%	21.5%	26.6%

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
Boynton Health Service	Above Average	2,216	98.5%	97.9%	98.9%
Catholic Charities Behavioral Health Clinic	Top Performer	172	100.0%	97.8%	100.0%
CCM Health	Top Performer	855	100.0%	99.6%	100.0%
Cedar Riverside People's Center	Above Average	162	84.6%	78.2%	89.3%
CentraCare Health	Above Average	10,759	93.7%	93.2%	94.1%
Children's Minnesota	Top Performer	217	100.0%	98.3%	100.0%
Cromwell Medical Clinic PLLC - IHN	Below Average	83	55.4%	44.7%	65.6%
Cuyuna Regional Medical Center	Above Average	681	93.2%	91.1%	94.9%
Dawson Clinic	Below Average	405	43.7%	39.0%	48.6%
Dr. VM Baich, PA	Average	32	81.3%	64.7%	91.1%
Duluth Family Medicine Clinic	Below Average	358	61.5%	56.3%	66.3%
Edina Sports Health & Wellness	Below Average	529	42.2%	38.0%	46.4%
Entira Family Clinics	Average	3,578	78.9%	77.5%	80.2%
Essentia Health	Below Average	15,090	65.5%	64.7%	66.3%
Fairview Health Services	Above Average	22,452	79.9%	79.4%	80.4%
Fairview Mesaba Clinics	Above Average	1,243	83.3%	81.1%	85.2%
Family Practice Medical Center of Willmar	Above Average	214	97.2%	94.0%	98.7%
France Avenue Family Physicians	Above Average	867	97.0%	95.6%	97.9%
Glencoe Regional Health Services	Above Average	616	95.8%	93.9%	97.1%
Glenwood Medical Center	Below Average	948	13.8%	11.8%	16.2%
Grand Itasca Clinic	Above Average	742	83.2%	80.3%	85.7%
Granite Falls Municipal Hospital	Below Average	287	20.9%	16.6%	26.0%
Gundersen Health System	Above Average	412	84.7%	80.9%	87.9%
Gundersen Saint Elizabeth's	Top Performer	169	100.0%	97.8%	100.0%
HealthPartners Central Minnesota	Above Average	862	89.4%	87.2%	91.3%
Health Partners Clinics	Above Average	12,718	89.9%	89.4%	90.4%

* Cell intentionally left empty

Ranking of Clinics Alphabetical with Health Score Ranking

https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf page 5

The image above is a portion of a report the demonstrates the public reporting and display the Health Score Rankings for each medical group for this measure for adults.

Adolescent Depression- PHQ-9/PHQ-9M Utilization

Partial Alphabetical Listing of Minnesota Medical Groups with Ranking of Results

Full list in report at

https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
STATEWIDE AVERAGE	*	19,574	78.4%	*	*
Allina Health	Above Average	1,751	92.9%	91.6%	94.0%
Alomere Health	Above Average	86	98.8%	93.7%	99.8%
Altru Health System	Below Average	321	31.5%	26.6%	36.7%
Amery Hospital and Clinic	Above Average	47	93.6%	82.8%	97.8%
Associated Clinic of Psychology	Average	131	84.7%	77.6%	89.9%

Medical Group Name	Health Score	Denominator	Rate	95% CI Lower	95% CI Upper
CentraCare Health	Above Average	1,171	96.4%	95.2%	97.3%
Children's Minnesota	Top Performer	532	100.0%	99.3%	100.0%
Cuyuna Regional Medical Center	Above Average	43	93.0%	81.4%	97.6%
Eagan Valley Pediatrics	Top Performer	53	100.0%	93.2%	100.0%
Entira Family Clinics	Above Average	139	89.9%	83.8%	93.9%
Essentia Health	Above Average	1,290	84.3%	82.3%	86.2%
Fairview Health Services	Above Average	848	87.6%	85.2%	89.7%
Fairview Mesaba Clinics	Above Average	63	88.9%	78.8%	94.5%
Grand Itasca Clinic	Above Average	48	93.8%	83.2%	97.9%
HealthPartners Central Minnesota	Above Average	61	96.7%	88.8%	99.1%
Health Partners Clinics	Above Average	593	91.9%	89.4%	93.8%
Hennepin Healthcare	Above Average	715	99.4%	98.6%	99.8%
Hutchinson Health	Below Average	87	67.8%	57.4%	76.7%
Lake Region Healthcare	Above Average	131	91.6%	85.6%	95.2%
Lakeland Mental Health	Below Average	295	57.3%	51.6%	62.8%
Lakewood Health System	Average	118	81.4%	73.4%	87.4%
Mankato Clinic	Above Average	342	96.5%	94.0%	98.0%
Mayo Clinic	Average	869	78.7%	75.9%	81.3%
Mayo Clinic Health System	Above Average	784	86.5%	83.9%	88.7%

* Cell intentionally left empty
Ranking of Clinics Alphabetical with Health Score Ranking
https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020_DepressionCare_Adults&Adolescents_Appendix.pdf page 9
The image above is a portion of a report the demonstrates the public reporting and display the Health Score Rankings for each medical group for this measure for adolescents.

Rates for this measure are also published annually on the consumer facing public website MN HealthScores https://www.mnhealthscores.org/ however, due to redesign of this depression measure to incorporate adolescents and a currently in progress website redesign, it is not available on MN HealthScores.

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]
Public reporting
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)
[Response Ends]

4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.

For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?

[Response Begins]
NA
[Response Ends]

4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.

A 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.

[Response Begins]
[Response Ends]

4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

Detail 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.

[Response Begins]
Performance results are provided to those being measured: (all primary care and psychiatry setting clinics in MN) in all of the following modalities/processes on an annual basis:
* Preliminary rate displayed to the practice immediately after file upload to the MNCM Data Portal
* Practices receive a password protected email that allows them to see their rates along with all other reporting practices prior to publication on MN HealthScores. There is a two week review process in which practices can identify issues or concerns with their data which can either be resolved or formal appeal submitted. (detail below)
* Practices receive actual and expected (risk adjusted) outcome rates and rating (top, above average, average and below average) prior to publication on the MNHealthScores.
* Data is published and updated on an annual basis on our consumer-facing website MNHealthscores at www.mnhealthscores.org
* Hard-copy reports are also provided on an annual basis (Health Care Quality Report, Health Care Disparities Report) https://mncm.org/reports/#community-reports
Assistance is provided to all practices via our support@mncm.org email or by telephone helpline at 612-746-4522.

[Response Ends]

4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

[Response Begins]
Process described in 4a.05
Educational services provided through Data Submission Resources on the MNCM corporate website at https://mncm.org/data-submission-resources/. Resources include the MNCM Academy which contains recorded classes in data submission processes and measure specific courses which include the Depression Care measures. Additionally, MNCM maintains a Knowledge Base and Help Desk support where users can submit questions or suggestions for measures, answers are recorded, categorized and can be accessed by all.

[Response Ends]

4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.

[Response Begins]
MNCM convenes the Measure Review Committee, a subcommittee of the Measurement and Reporting Committee to review all measures on a rotating cycle. This multi-stakeholder committee consisting of representatives from providers, health plans, consumers and purchasers of healthcare, reviews each measure publicly reported by MNCM to determine measure value, ongoing opportunity for improvement and variability among practices as well as feasibility and burden. Committee members complete a structured preliminary survey of each measure and results and comments are tabulated

and committee discussion occurs culminating in a vote for: continue/ refer for ad-hoc review or redesign/ transition to monitoring or retiring the measure.
Feedback from users is also obtained in the following ways:
* Questions and comments coming through our support email and telephone hotlines at MNCM
* Public comment email publiccomment@mncm.org
* Formal annual public comment process in collaboration with the MN Department of Health
* Questions and comments from national use in federal programs (QNet and JIRA help-desks)
* NCQA's adaptation of the depression remission, response and utilization of the PHQ-9 measures into the HEDIS program, Electronic Clinical Data Systems (ECDS) methodology.

[Response Ends]

4a.08. Summarize the feedback obtained from those being measured.

[Response Begins]
Measure Review Committee and many medical groups identify challenges with the technical replication of this measure in the medical group's internal systems (index event and follow-up window and the difficulty in maintaining ongoing contact with patients who are depressed.
Periodically, MNCM surveys all medical groups in MN to assess value in measures and feasibility/ ease/ difficulty in data collection and submission to MNCM for measure rate calculation. Ease/ difficulty ratings for the depression measures improved by 9% and we expect significantly less burden with the transition to the new data collection methodology PIPE. 57% of those surveyed rated the depression measures as high/moderate value.

[Response Ends]

4a.09. Summarize the feedback obtained from other users.

[Response Begins]
2015 at the request of the measure review committee, implemented technical change to MNCM data portal programming for the re-indexing of patients following the held assessment period.
2016 request for the consideration of incorporating adolescents into the current measure construct, increasing the follow-up window, use of additional PRO tools and review of exclusions.
[Response Ends]

4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

[Response Begins]
For 2020 Report Year (dates of index event 1/1/2018 to 12/31/2018)
1. Incorporate adolescents into the depression measures
* Modify age range to include adolescents; age 12 and older
* Report measures as two separate stratifications by age (not combined); ages 12 to 17 and ages 18 and older
Reason: Depression is a significant problem for adolescents, affecting an estimated 11% of the population. Many mental health conditions are evident by age 14 and the consequences of adolescent depression can have lifelong impact.
2. Widen the follow-up assessment window to +/- 60 days for all populations and all response and remission measures
* Six month measures assessment window expands from 5 to 7 months to 4 to 8 months
* Twelve month measures assessment window expands from 11 to 13 months to 10 to 14 months
Reason: Allowing a more reasonable assessment window that still fits the clinical course of recovery, allows for a comprehensive course of treatment and increases provider buy-in.
3. Patient Reported Outcome Tools for index/denominator and measuring outcomes of remission and response are the PHQ-9 and PHQ-9M
* Add the PHQ-9M as a PRO tool that can be used
* Providers may elect to use either tool; no measure construct restriction for age
Reason: 21 additional tools were reviewed against standardized criteria, very few had cut-points for severity levels of depression or remission. Potential threat to comparability was determined; using PRO tools with significantly different

numbers of questions could impact the response measures (50% or greater in improvement of scores) in addition to denominator comparability.
4. Modifications to exclusions include the following:
* Personality disorders narrowed to emotionally labile conditions and moved to the allowable exclusion category
* Add exclusion value set for schizophrenia or psychotic disorder as a required exclusion
* Add exclusion value set for pervasive developmental disorder as an allowable exclusion
Reason: Recommendation from NQF behavioral steering committee to examine the personality disorder exclusion.
5. Remove denominator criteria for behavioral health settings that stipulates the diagnosis of major depression or dysthymia needs to be in the primary position.
* Relates to new exclusion for schizophrenia or psychotic disorder; no longer necessary
Reason: simplification of measure, behavioral health providers determine position order of diagnosis is irrelevant.
Please refer to either the data dictionary (attachment in **sp.11)** for the summary of redesign activities and changes to value sets or the electronic newsletter with links to details at http://mncm.org/?s=depression.

[Response Ends]

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, 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, provide an explanation. 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.

[Response Begins]
Measure does demonstrate a trend in improvement over time with opportunity for continued improvement and variability among medical groups as demonstrated in **1b.02)**.The denominator of eligible patients has doubled in the last 10 years (108,261 in 2010 to over 244,00 in 2020) demonstrating more patients are being screened for depression with appropriate recognition and diagnosis of the condition. This, in a large part is due to the integration of the PHQ-9 into practice workflows for both screening, diagnosis and then measuring outcomes.
[Response Ends]

4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.

[Response Begins]
No unintended negative consequences identified.
[Response Ends]

4b.03. Explain any unexpected benefits realized from implementation of this measure.

[Response Begins]
Increased screening for depression, diagnosis of major depression or dysthymia and increase in rates of follow-up assessments for the managing of successful outcomes of response and remission.

- Increasing widespread use of a simple but effective PRO tool that can be used for screening, diagnosis and the monitoring of treatment outcomes for depression
- Increased national use of the measure, adaptation of the measure for use by health plans (HEDIS)
- Incorporation of adolescents helps address a significant condition that can have lifelong impacts

[Response Ends]

Criteria 5: Related and 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.

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02, if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

(Can search and select measures.)
[Response Begins]
0710e: Depression Remission at Twelve Months
1885: Depression Response at Twelve Months- Progress Towards Remission
0711: Depression Remission at Six Months
1884: Depression Response at Six Months- Progress Towards Remission
[Response Ends]

5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus or target population).

(Can search and select measures.)
[Response Begins]
[Response Ends]

5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]
Several depression assessment measures have had NQF endorsement removed.
[Response Ends]

5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]
Yes
[Response Ends]

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

[Response Begins]
Measure definitions of these related measures, all developed and stewarded by MNCM are completely harmonized
[Response Ends]

5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

Provide analyses when possible.

[Response Begins]
There are related, complimentary measures for depression remission, response that are companion measures with this process measure. MN Community Measurement is the measure steward for these related measures and they are completely harmonized. The remission measures are considered the "gold standard" of depression outcomes and measure the same population of patients at two different points in time, six and twelve months after index contact with diagnosis and elevated PHQ-9. The response measures, also at six and twelve months are considered as progress towards the desired goal of remission with a reduction in PHQ-9 score of greater than 50% representing a reduction in the severity of symptoms.
[Response Ends]