NATIONAL QUALITY FORUM
Driving measurable health improvements together

MEASURE WORKSHEET

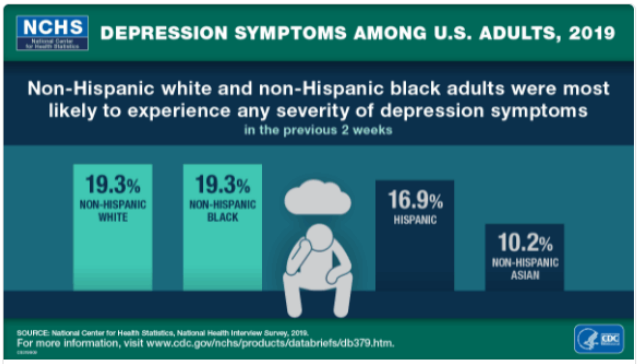
This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

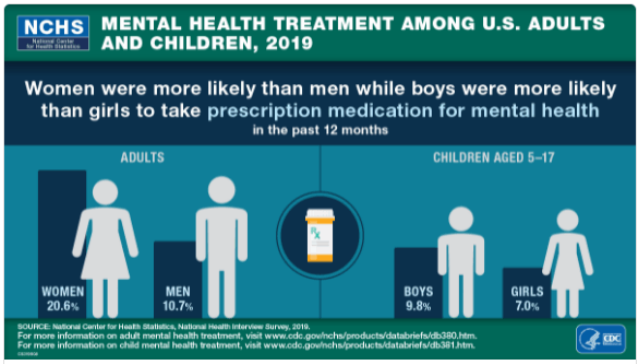
To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1885
Corresponding Measures:
Measure Title: Depression Response at Twelve Months- Progress Towards Remission
Measure Steward: MN Community Measurement
sp.02. 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 major depression or dysthymia who are progressing towards remission by achieving a response (PHQ-9 or PHQ-9M score reduced by 50% or greater) twelve months (+/- 60 days) after an index visit.
1b.01. Developer Rationale:
Adults:
Depression is a common and treatable mental disorder. The Centers for Disease Control and Prevention states that in 2019 (1)

- 2.8% of adults experienced severe symptoms of depression, 4.2% experienced moderate symptoms, and 11.5% experienced mild symptoms in the past 2 weeks.
- The percentage of adults who experienced any symptoms of depression was highest among those aged 18–29 (21.0%), followed by those aged 45–64 (18.4%) and 65 and over (18.4%), and lastly, by those aged 30–44 (16.8%).
- Women were more likely than men to experience mild, moderate, or severe symptoms of depression.
- Non-Hispanic Asian adults were least likely to experience mild, moderate, or severe symptoms of depression compared with Hispanic, non-Hispanic white, and non-Hispanic black adults.





Prevalence of Depression in Adults and Children; Centers for Disease Control 2019

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)

Adolescents:

- In 2019, 16% of the population ages 12–17 had at least one MDE during the past year, a higher prevalence than that reported in each year between 2004 (9%) and 2014 (11%).
- Among youth ages 12–17 in each year between 2004 and 2019, the prevalence of MDE was more than twice as high among females (ranging from 12% to 23%) as among males (ranging from 4% to 9%).
- The prevalence of MDE in 2019 was lowest among youth ages 12–13 (11%) compared with youth ages 14–15 (16%) and ages 16–17 (20%).
- Between 2004 and 2019, the prevalence of MDE increased for both genders among all three age groups (12–13, 14–15, and 16–17).
- The percentage of youth with MDE in the past year receiving treatment for depression increased between 2004 (40%) and 2019 (43%), but this increase was not statistically significant. Treatment was higher among females (46%) than among males (37%) in 2019. (5)
- In 2015, 9.7% of adolescents in MN who were screened for depression or other mental health conditions, screened positively.

---

sp.12. Numerator Statement: The number of patients in the denominator who achieved a response as demonstrated by a PHQ-9 or PHQ-9M score reduced by 50% or greater twelve months (+/- 60 days) after an index visit.

sp.14. Denominator Statement: Adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) with major depression or dysthymia and an initial (index) PHQ-9 or PHQ-9M score greater than nine.

sp.16. 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: Outcome: PRO-PM

sp.28. Data Source: Electronic Health Records

sp.07. Level of Analysis: Clinician: Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: 03/04/2014

Most Recent Endorsement Date: 03/04/2014

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 ***health outcome*** measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

The developer provides the following description for this measure:

- This is a maintenance patient-reported outcome performance measure (PRO-PM) at the Clinician: Group/Practice level of analysis that assesses the percentage of adolescent patients (12 to 17 years of age) and adult patients (18 years of age or older) with major depression or dysthymia who are progressing towards remission by achieving a response PHQ-9 or PHQ-9M score reduced by 50 percent or greater twelve months (+/- 60 days) after an index visit.
- The developer provides a logic model that depicts the assessment of major depressive disorder (MDD) or dysthymia using the PHQ-9/M, which leads to treatment and/or therapy, leading to continued monitoring and if needed a step-wise approach to treatment, and finally the response at twelve months (+/- 60 days). The ideal response is a decline in PHQ-9/M score of >=50 percent from baseline.

Summary of prior review in 2014

- The Standing Committee stated that strong evidence was presented to support the measure focus, and practice groups adopting the measure markedly improved their screening and response rates.

Changes to evidence from last review

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

☒ The developer provided updated evidence for this measure:

- For follow up recommendations, the developer cites the Institute for Clinical Systems Improvement Health Care Guideline *Depression in Primary Care*, which concludes that clinicians should establish and maintain follow-up with patients. The quality of the evidence for this guideline is low and the strength of the recommendation is strong. The guideline states that the accountable entity can influence the outcome through proactive follow up by phone and in person in conjunction with behavioral health specialists to provide psychotherapy and/or pharmacotherapy treatments and monitor progress through regular administration of the PHQ-9.
    - The developer uses guidance from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) to assess severity of PHQ-9 score with cut points of 10-14, 15-19, >20.
- The literature shows that the PHQ-9 is an effective management tool and should be used routinely during follow-up 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).
- The developer states that a five-point drop in PHQ-9 score is considered the minimal clinically significant difference (Trivedi, 2009).

- The developer states the recommendation that a clinician should establish and maintain follow-up with patients with high PHQ-9 scores because appropriate, reliable follow-up is highly correlated with improved response and remission scores, as well as with the improved safety and efficacy of medications, and helps prevent relapse.
    - Proactive follow-up via in person or telephone, based on the collaborative care model, has been shown to significantly lower depression severity (Unützer, 2002).
    - In clinical effectiveness trials conducted in clinical practice settings, 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. (Hunkeler, 2000; Simon, 2000).
- The developer cites the Veterans Affairs Department of Defense *Clinical Practice Guidelines for Depression* for the treatment algorithm related to major depressive disorder (MDD) and persistent depressive disorder. The algorithm outlines identification of depression through the PHQ-2 followed by assessment and triage where MDD is identified, then management in which the patient undergoes treatment and achieves remission. The algorithm details additional considerations for treatment of both mild/moderate MDD (such as select monotherapy or combination therapy (pharmacotherapy/psychotherapy)) and severe MDD (such as refer to specialty level care).
- As evidence that the target population values the measured outcome and finds it meaningful, the developer cites a qualitative study though which patient feedback on relevant treatment outcomes in depression was collected (Kan et al, 2020). They found that the majority of patients had goals related to regaining daily activities and social functioning, while those with chronic depression stressed the need to find new ways of functioning, even if they are not able to return to full social functioning.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *Does the evidence support the time period for measurement and degree of improvement required to meet the measure?*

Guidance from the Evidence Algorithm

Measure assesses performance on a patient reported health outcome (Box 1) -> the relationship between the measured/patient reported health outcome and at least one healthcare action is demonstrated by empirical data (Box 2) -> Rate as PASS

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

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 2019, among 120,344 adults from 550 clinics, 17 percent (range 0-32.7%) of patients with an index event had a depression response (>=50% from baseline) at 12 months. Rate of follow up at 12 months decreased from 41.8 percent in 2019 to 39.6 percent in 2020, likely influenced by COVID-19.

- In 2019, among 11,658 adolescents (ages 12-17) from 118 clinics, 14.5 percent (range 0-29.1%) of patients with an index event had a depression response (>=50% from baseline) at 12 months. Rate of follow up at 12 months decreased from 38.9 percent in 2019 to 35.6 percent in 2020, likely influenced by COVID-19.

Disparities

- In 2019, among 120,344 adults from 550 clinics, there was a differential response based on insurance status (13.1%-MCO v 17.7%-Other insurers) and race. Outcomes based on MCO status and race ranged from a high of 14.1 percent among white patients to 9.6 percent among black patients. Outcomes based on other insurance status and race ranged from a high of 18.1 percent for white patients to 7.8 percent for patient race not reported and 10.1 percent for black patients.
- In 2019, among 11,658 adolescents (ages 12-17) from 118 clinics, there was a differential response based on insurance status (14%-MCO v 16%-Other insurers) and race. Outcomes based on MCO status and race ranged from a high of 14 percent among white patients to 7 percent among black patients. Outcomes based on other insurance status and race ranged from a high of 17.8 percent for multi-race patients to 4.8 percent for patient race not reported, 12.3 percent for black patients, and 15.1 percent for white patients.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

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

Committee Pre-evaluation Comments:

1a. Evidence

- Yes. Strong empirical evidence when measure was established. Multiple actions the provider can do to achieve a change in measure results.
- This measures actual response to treatment using PHQ9 and is an important indicator of treatment outcome
- The developer provides data that use of the outcome measure can detect depression and lead to the development of treatment plans and improved clinical outcomes/remission. The application is direct. The process leads to detection and diagnosis, treatment, and improved outcomes. They cite the same evidence around the patients having social and daily functioning goals.
- Maintenance outcome measure-addition of clinical literature and a qualitative study showed the relationship between the measured/patient reported health outcome and at least one healthcare action is demonstrated by empirical data.
- Evidence is good

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

- Do appear to be gaps in improvement - 17% of adults and 14.5% of adolescents had depression response at 12 months. Decrease during COVID-19 also illustrative of performance gap and urgency. Disparities documented by race, insurance status for both adults and adolescents.
- Depression increased during COVID pandemic, so measuring improvement in depression is now more important than before.

- Reassessment rates decreased at 12 months from 2019 to 2020 attributed to covid-19. Disparities for insurance status and race.
- A performance gap was acknowledged pertaining to responses. Yes measurement was provided. Disparities were identified by subgroup. Managed care/Medicaid patients who were black or adolescent had much lower outcomes.
- The performance gap is real and demonstrates disparities

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 for the instrument-based measure also include the specific instrument (e.g., PROM(s)); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and calculation of response rates to be reported with the performance measure results.
- The developer states that several updates have been made to the measure specifications including: incorporating adolescents ages 12 to 17, addition of the PHQ-9M (modified for teens) PRO tool, expansion of the assessment window to +/- 60 days, modification of exclusion value set for personality disorder, addition of exclusions for schizophrenia and pervasive developmental disorder, and removal of the requirement that the depression diagnosis be in the primary position for behavioral specialty.

Reliability Testing:

- Reliability of PHQ-9:
    - The developer refers to prior evidence of encounter-level reliability of the PHQ-9 from the literature.
        - Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study and
        - Cronbach's alpha of 0.86 in the PHQ OBGYN Study.
        - Test-retest showed a correlation of 0.84 between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours.
    - The developer describes differences between PHQ-9 and PHQ-9M, which they characterize as "slight." The PHQ-9 has been tested in adolescents. The PHQ-9M has not been tested separately. However, the developer asserts that this is not necessary, given the minor differences between the questionnaires

- Reliability testing conducted at the Accountable Entity Level
    - The denominator identification period (index) for the testing data was 11/1/2017 to 10/31/2018. The measure assessment period was through 12/30/2019; reported in 2020.
        - Measure Assessment Period: For each patient, the measure assessment period begins with an index event and is 14 months (12 months +/- 60 days) in length. The assessment period is held constant to assess the same denominator of eligible patients for outcomes of remission and response at both six and twelve months.
    - Over 115 medical groups representing 788 clinics were included in the testing of this measure, representing 118,132 adults and 7,237 adolescents. Reliability statistics were provided at the clinic level for all clinics with >=30 patients in the denominator.
        - The developer used a beta binomial test to assess reliability.
            - For adults, signal-to-noise was 0.92 (550 clinics, 118,132 patients). A graph shows the range of values but individual data points are not provided.
            - For adolescents, the performance score was 0.84 (118 clinics, 7,327 patients). A graph shows the range of values but individual data points are not provided.
            - The developer states that with a reliability score exceeding 0.91 and 0.84, there is the ability to distinguish higher performing clinics from lower performing clinics for both adults and adolescents.
    - Reliability scores increase slightly from the 2013 submission among the adult population. The previous submission (2013) demonstrated reliability of 0.88 for adults, using a beta binomial test.
- The developer states that a signal-to-noise score of greater than 0.70 indicates that it is acceptable to draw conclusions using this data.

Questions for the Committee regarding reliability:

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

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 testing conducted at the Patient/Encounter Level:
    - Validity of 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 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 percent of groups passed with no errors; 58 percent of those that submitted data passed initial quality checks; 30 percent of groups that submitted data were audited; 94 percent passed the audit.
    - Percent agreement statistics or positive and negative predictive values were not provided.
    - The developer does not provide additional validity testing of the PHQ-9M.
    - The developer does not present results for all critical data elements; therefore, this testing does not meet the NQF threshold for critical data element testing.
- Validity testing conducted at the Accountable Entity Level:
    - The developer presents empirical testing at the accountable entity level, testing against several different constructs.
    - Correlation between depression remission (PHQ-9 < 5) at six months and depression response (PHQ-9 greater than or equal to 50% improved from index initial PHQ-9 score) at six months. The developer hypothesizes that clinics that have high response rates are also likely to have high remission rates for both adults and adolescents.
        - R-squared (adults) = 0.9051
        - R-squared (adolescents) = 0.7896
    - Correlation between depression response at six months and rates of follow-up with a PHQ-9/9M at six months. The developer hypothesizes that patients who receive regular screening are more likely to achieve remission for both adults and adolescents.
        - R-squared (adults) = 0.7967
        - R-squared (adolescents) = 0.7924
    - Correlation between patients who achieve remission at six months and patients who achieve response at six months but not remission. The developer hypothesizes that clinics that have high response rates are also likely to have low response with no remission rates for both adults and adolescents.
        - R-squared (adults) = 0.3578
        - R-squared (adolescents) 0.2366
    - Correlation between patients with depression outcome and diabetes outcome. The developer hypothesizes that there will be a weak but positive correlation between these two chronic conditions for adults only.
        - R-squared (adults) = 0.1406

Exclusions

- Exclusions for this measure include:
    - Bipolar or personality disorder (n=1,122)- updated in 2020
    - Schizophrenia or psychotic disorder (n=1,606)- new in 2020

- Pervasive developmental disorder (n=909)- new in 2020
- Permanent resident of a nursing home (n=104)
- Enrolled in hospice (n=137)
- Patients who die (n=423)

- By applying the exclusions, 3.45 percent of the patient population (4,834 patients) were excluded from the measure.
- The developer lists bipolar diagnosis and active Schizophrenia, or Psychotic Disorder as required exclusions and the remaining exclusions as allowable. The developer states that because this is a longitudinal measure the allowable exclusion may occur during the course of the measurement period.

Risk-Adjustment

- The measure is risk adjusted using a logistic regression model to create an indirect standardization risk adjustment (expected value). Performance is measured against the expected value for the given case mix of the clinic. Separate models were run for adults and adolescents.
- Risk variables included in the model include age, initial PHQ-9/PHQ-9M score, insurance product and patient neighborhood deprivation index (based on zip code). Deprivation index is new in 2021. Deprivation index includes use of SNAP benefits, living under the poverty level, unemployed status, public assistance, and single female with children.
- The developer considered race, ethnicity, language and country of origin variables for the model. They did have an impact on the score, but the developer could not prove both sufficient conceptual basis for their inclusion and that they were not confounding factors. The developer also thought their application introduced the potential for implicit bias. The social deprivation index was included as a proxy for these social determinants of health with the decision that geography/neighborhoods are what matter.
- The developer provided the model estimates but model discrimination statistics were not included. The developer states that all variables have a Chi-squared p value of less than 0.0001, but a C-statistic and other model fit or calibration statistics were not provided.

Meaningful Differences

- Variability of rates among medical groups around the statewide average was as follows:
  - Adults: 17.0% (range 0% to 37.2%), using 120,344 patients from 550 clinics
  - Adolescents: 14.5% (range 0% to 29.1%), using 11,658 patients from 118 clinics
- The developer reports that twelve month follow-up has the widest variation among medical groups, and that overall rates are low.
- The developer does not describe the statistical methods for identifying meaningful differences.
- The developer provided information in 2b.28 on the risk adjusted results that show differences. In the adult model, 85 of the 550 groups/practices performed above expectations, 106 groups/practices performed below and 359 performed as expected. In the adolescent model, 111 of 118 groups/practices performed as expected, while one performed below expectations and one performed above expectations.

Missing Data

- The developer states that MN has made incremental improvements in rates of follow-up PHQ-9 at 12 months, from 17.0% in 2010 to 41.8% in 2019 for adults. Adolescents, a new population for this measure have a 2019 follow-up rate of 38.9%.
- The developer states that missing data (follow-up PHQ-9 patient reported outcome assessment) is not an issue as those patients who are not re-assessed in follow-up remain in the denominator and are treated as if they are not in remission, but that low outcome rates are not solely attributed to lack of follow-up. A portion of patients are still experiencing symptoms of depression and are not in remission. A separate analysis for patients who were assessed with a follow-up PHQ-9 demonstrates that remission was at 24 percent while significant depression symptoms persisted for 49 percent of the patients (24% moderate, 15% major, and 10% severe).
    - There is a companion related measure that allows medical groups to understand their use of the PHQ-9/ PHQ-9M tool, NQF # 0712 Depression Utilization of PHQ-9M (also under maintenance review this cycle). This measure reports the rate of tool administration for patients with a diagnosis of depression or dysthymia seen during a four month
- Missing follow up data is included in the denominator and patients who are not re-assessed are treated as if they are not in remission.

Comparability

- The measure only uses one set of specifications for this measure.

Questions for the Committee regarding validity:

- *Do you have any concerns regarding the validity of the measure (e.g., exclusions, risk-adjustment approach, etc.)?*
    - *Was is appropriate to compare the current measure using the PHQ-9 with other measures that use the same tool?*
    - *Do you have any concerns about the way missing data is categorized in this measure?*
- *Is any additional data needed on data element testing?*
- *Is additional information needed on the risk adjustment model and model performance?*

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

Committee Pre-evaluation Comments:

2a. Reliability

- 2a1. Reliability-Specifications
    - No concerns regarding reliability specifications. PHQ-9M appears to minimally different from PHQ-9 (just translating same questions to younger audience) so not concerned about lack of separate reliability testing. Measure specifications seem adequate.
    - Data systems may not accurately capture change in PHQ9 score. May be difficult for a provider to report.
    - Data elements clearly defined and descriptors provided. All steps are clear. No concerns about measure being implemented consistently.
    - No concerns, clearly defined.
    - This is fine

- 2a2. Reliability – Testing
    - No
    - PHQ-9M should undergo reliability testing, although it is doubtful it will differ from PHQ9 so not as much of a concern.
    - No.
    - No concerns.
    - No concerns

2b. Validity

- No.
- No
- Does not pass because of missing critical data elements.
- Concern-The developer does not present results for all critical data elements; therefore, this testing does not meet the NQF threshold for critical data element testing.
- No

2b2-2b6. Potential threats to validity

- 2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment)
    - Exclusions are clinically appropriate. Risk adjustment handled appropriately, with appropriate conceptual relationship between social risk factor variables and measure focus, and appropriate design and testing.
    - Risk adjustment is warranted due to disparities in depression care. Exclusions are appropriate.
    - Yes, there is a conceptual relationship between potential social risk factor variables and measure focus. Variables present at the start of care. Yes to all items.
    - Results are acceptable.
    - This is adequate
- 2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data)
    - Would be good to understand any statistical methods used for identifying meaningful differences. Same concerns are previously that missing data is included in the denominator and patients who are not re-assessed are treated as if they are not in remission.
    - Possibly. If there was no follow-up data vs. the original PHQ9 score the person should be excluded from the denominator so as to not falsely lower results.
    - The developer includes differences in performance among clinical sites. Reports 12 month follow up to be the lowest. Reports on differences among clinical sites. Only one set of specifications. Follow up rates are low for six-months and those patients will not be included in the numerator.
    - I do not believe the missing data constitutes a threat to the validity.
    - No problems in this area

Criterion 3. Feasibility

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

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

- The measure is captured in electronic health records. Groups can successfully extract the tool information from their EHR.
- MNCM developed a direct data submission process, whereby medical groups submit a patient level data file for rate calculation and public reporting. MNCM is implementing a new data collection method that serves as a warehouse of clinical data where measures are calculated centrally. No fees are associated with this program.

Questions for the Committee:

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

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

Committee Pre-evaluation Comments:

3. Feasibility

- No concerns with EHR data usage and extraction approach.
- May be difficult for some providers to record results of depression screening and adequately report. Also, a 6 month timeframe may be too short based on when patients receive follow-up care (vs. medication management only). Would need to figure out how to implement at a national level.
- All of the data elements are routinely generated and available in electronic format. They are implementing a new data collection method where measures are calculated centrally. I do not have concerns around the data collection strategy and its operational use.
- No concerns.
- 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

- Performance results are provided to all medical groups who submit data for this state-wide measure. Results are provided annually.
- Measure is publicly reported on the MN HealthScores website and as a part of the MNCM Annual Health Care Quality Report, Annual Disparities by Insurance Type and Disparities by Race, Ethnicity, Language, Country of Origin, and is the focus of several issue briefs.
- The measure is used in all primary care clinics in MN and bordering communities in Wisconsin, North Dakota, South Dakota and Iowa.
- The measure was selected for inclusion as a quality metric for CMS' CMMI Innovation Model Kidney Care First.

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.
- Although this measure is not used in a CMS program, the developer uses a CMS JIRA ticketing system to collect and respond to questions about other measures with which NQF #1885 is harmonized. The developer has made improvements to measure NQF #1885 in response feedback received via CMS JIRA tickets as well.
- In response to feedback, the expert panel to review updates to the measure specifications. Based on feedback received from the developer-convened multi-stakeholder expert workgroup, the developer made the following changes to the measure: addition of the adolescent population, widening the follow-up assessment window, 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.

***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

- Prior to specifications changes, follow up at 12 months improved from 17 percent in 2010 to 41.8 percent in 2019 for adults. Adolescents have a 2019 follow up rate of 38.9 percent.
- Due to recent redesign of the measure, the developer is not able to provide trend data on the measure as specified.

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.
- The developer conducted a survey of medical groups in MN. The developer found that 55.6 percent of medical groups rated the measure as moderate or high value.

Potential harms

- The developer does not note potential harms of the measure.

***Questions for the Committee**:*

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

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

Committee Pre-evaluation Comments:

4a. Use

- Measure is publicly reported and used in current and future accountability programs. Developed provides results to entities and facilitates bi-directional feedback.
- Yes
- Performance results are reported back annually to groups who submit data. Publicly reported on the MN Health Scores. Used as quality metric for CMS CMMI Innovation Model Kidney First Care. Groups being measured have been given their results and they are provided the opportunity to appeal. Feedback has been gathered by an expert panel, and changes were implemented based on the feedback.
- Opportunity for feedback was provided and the developer incorporated the feedback in the measure.
- It's widely publicly reported, feedback has been broad and resulted in useful measure modifications

4b. Usability

- Promising improvement in rates from 2010 to 2019 for adults. No apparent harms or unintended consequences. Good to see positive response from MN medical group community around the measure.
- Benefits outweigh the risks.

- Tracking scores over time can assist in the development and use of effective interventions for treatment of depression. The feedback to specific sites allows them the opportunity for quality improvement of patient care. 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.
- No harms. The benefit is to continue to promote quality care.
- No harms and it is widely used and improvement is occurring albeit slowly

Criterion 5: Related and Competing Measures

Related measures

- NQF #1884 Depression Response at Six Months- Progress Towards Remission
- NQF #0712 Depression Assessment with PHQ-9/ PHQ-9M
- NQF #0710e Depression Remission at Twelve Months
- NQF #0711 Depression Remission at Six Months

Harmonization

- The developer attests that the related measures are harmonized to the extent possible.

Committee Pre-evaluation Comments:

5: Related and Competing Measures

- 4 related measures which seem, logically, to be harmonized to the extent possible.
- Several related measures. Suggest combining the 6 and 12 month response measures (but not remission) into a single measure.
- All related measures have been harmonized.
- No competing measures; NQF #1884 Depression Response at Six Months- Progress Towards Remission • NQF #0712 Depression Assessment with PHQ-9/ PHQ-9M • NQF #0710e Depression Remission at Twelve Months • NQF #0711 Depression Remission at Six Months
- No competing measure and it's well harmonized with the suite of 4 other MNCM measures

Public and NQF Member Comments (Submitted as of June 15, 2022)

Member Expression of Support

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

Comments

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

Hello, During the process of submitting our scientific testing for this measure NQF# 1885 Depression Response at Twelve Months, we inadvertently did not include the c-statistic for this measure. This statistic was calculated during the logistic regression procedure but the clinical staff completing the application did not recognize the c-statistic in part due to the large number of pairs and the spacing of the table. The calculated concordance (c-statistic) for this measure was 0.587 (adults) and 0.556 (adolescents) which meet the criteria for a well calibrated model. Association of Predicted Probabilities and Observed Responses Adults Percent Concordant 58.7Somers' D 0.173 Percent Discordant

41.3Gamma 0.173 Percent Tied 0.0Tau-a 0.049 Pairs 2042832300c 0.587 Association of Predicted Probabilities and Observed Responses Adolescents Percent Concordant 55.6Somers' D 0.113 Percent Discordant 44.4Gamma 0.113 Percent Tied 0.0Tau-a 0.028 Pairs 16879016c 0.556 Please consider this additional information in the standing committee's assessment of the risk adjustment model. 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 express our concerns on the evidence and testing provided in support of this measure. While the AMA agrees that it is useful to understand the rate of response for individuals diagnosed with depression, we do not believe that the developer provided sufficient evidence demonstrating that depression scores can be successfully reduced by at least 50% across the defined patient population within a twelve-month timeframe nor was any evidence provided supporting this requirement of 50%. For example, would the measure better capture clinical care and patient outcomes if it measured a minimal clinically significant difference in the depression score. It is important that the data demonstrate that practices can implement structures or processes that lead to improved outcomes and the measure results in rates that truly reflect the quality of care delivered by a practice rather than differences in patient mix or other factors outside of the practice's control. We also 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 that the gaps in evidence and 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.

Scientific Acceptability Evaluation

RELIABILITY: SPECIFICATIONS

1. **Have measure specifications changed since the last review? ☒ Yes ☐ No**
2. **Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? ☒ Yes ☐ No**
3. **Briefly summarize any changes to the measure specifications and/or concerns about the measure specifications.**
   - The developer states that several updates have been made to the measure specifications including: incorporating adolescents ages 12 to 17, addition of the PHQ-9M (modified for teens) PRO tool, expansion of the assessment window to +/- 60 days, modification of exclusion value set for personality disorder, addition of exclusions for schizophrenia and pervasive developmental disorder, and removal of the requirement that the depression diagnosis be in the primary position for behavioral specialty.
   - Measure specifications for the instrument-based measure also include the specific instrument (e.g., PROM(s)); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and calculation of response rates to be reported with the performance measure results.

RELIABILITY: TESTING

4. **Did the developer conduct new reliability testing? ☒ Yes ☐ No**

   4a. If no, summarize the Standing Committee's previous feedback:

4b. If yes, describe any differences between the new and old testing and summarize any relevant Standing Committee's feedback from the previous review:

5. **Reliability testing level: ☒ Accountable-Entity Level ☒ Patient/Encounter Level ☐ Neither**
6. **Reliability testing was conducted with the data source and level of analysis indicated for this measure:**

   ☒ Yes ☐ No
7. If accountable-entity level and/or patient/encounter level reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of patient-level data conducted?

   ☐ Yes ☐ No
8. **Assess the method(s) used for reliability testing:**
   - The developer refers to prior evidence of encounter-level reliability of the PHQ-9 from the literature
   - The developer describes differences between PHQ-9 and PHQ-9M, which they characterize as "slight." The PHQ-9 has been tested in adolescents. The PHQ-9M has not been tested separately. However, the developer asserts that this is not necessary, given the minor differences between the questionnaires.
   - The developer presents empirical testing at the accountable entity level using a beta-binomial model.
9. **Assess the results of reliability testing**
   - Reliability of PHQ-9:
     - The developer refers to prior evidence of encounter-level reliability of the PHQ-9 from the literature.
     - Cronbach's alpha of 0.89 in the PHQ-9 Primary Care Study and
     - Cronbach's alpha of 0.86 in the PHQ OBGYN Study.
     - Test-retest showed a correlation of 0.84 between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours.
   - Reliability testing conducted at the Accountable Entity Level
     - The developer used a beta binomial test to assess reliability. For adults, signal-to-noise was 0.92 (550 clinics, 118,132 patients).
     - For adolescents, the performance score was 0.84 (118 clinics, 7,327 patients).
     - The developer states that with a reliability score exceeding 0.91 and 0.84, there is the ability to distinguish higher performing clinics from lower performing clinics for both adults and adolescents.
     - The developer states that a signal-to-noise score of greater than 0.70 indicates that it is acceptable to draw conclusions using this data.
10. 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.

    ☒ Yes ☐ No ☐ Not applicable
11. Was the method described and appropriate for assessing the reliability of ALL critical data elements?

    ☒ Yes ☐ No ☐ Not applicable (patient/encounter level testing was not performed)
12. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and all testing results):

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

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity 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)

13. **Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.**
    - The specifications are precise, unambiguous, and complete (box 1) → Empirical reliability testing was conducted on the measure at the appropriate levels at both the encounter and the accountable entity levels (box 2) → Reliability testing was conducted with computed performance measure scores (box 4) → Method described was appropriate for assessing proportion of variability (box 5) → There is a moderate certainty that the performance scores are reliable (box 6a) → Rate at MODERATE

VALIDITY: TESTING

14. **Did the developer conduct new validity testing? ☒ Yes ☐ No**

    14a. If no, summarize the Standing Committee's previous feedback:

    14b. If yes, describe any differences between the new and old testing and summarize any relevant Standing Committee's feedback from the previous review:

15. **Validity testing level (check all that apply):**

    ☐ Accountable-Entity Level ☐ Patient or Encounter-Level ☒ Both

    NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

16. **If patient/encounter level validity testing was provided, was the method described and appropriate for assessing the accuracy of ALL critical data elements? NOTE:** Data element validation from the literature is acceptable.

    ☒ **Yes**

    ☐ **No**

    ☐ **Not applicable** (patient/encounter level testing was not performed)

17. **Method of establishing validity at the accountable-entity level:**

    ☐ **Face validity**

    ☒ **Empirical validity testing at the accountable-entity level**

    ☐ **N/A (accountable-entity level testing not conducted)**

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

    ☒ **Yes**

    ☐ **No**

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

19. **Assess the method(s) for establishing validity**
    - The developer references testing of construct validity in the literature, using mental health professional re-interview as the criterion standard.
    - The developer conducted data element and performance score validity.
    - For data element validity, the submitted data was authenticated via the direct data submission process which included denominator certification, data quality checks, validation audit, and a two-week medical group review period.

- Performance score validity was performed by comparing this measure to a similar measure. The developer hypothesized that clinics that do well achieving a response (PHQ-9 > 50 percent improved from index initial PHQ-9 score) would also do well in achieving remission (PHQ-9 < 5).

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

- For validity of the PROM, PHQ-9, the developer found 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, and for ROC analysis, the area under the curve for the PHQ-9 in diagnosing major depression was 0.95
- The developer reports that for pre-submission and post-submission, 49% and 58% passed data quality checks respectively. During the audit of data source, 30% of groups that submitted data were audited; of which 94% passed the audit. The developer concludes that there was high compliance with critical data element validity as demonstrated by annual validation audit processes.
- The developer reports that the adult stratification demonstrates a high correlation [R squared 0.9051] against a similar measure and the adolescent stratification demonstrates a lower correlation value [R squared 0.6026].
- The developer reports that if the coefficient value lies between ± 0.50 and ± 1, then there is said to be a strong correlation.
    - The developer states that the adolescent stratification may have a lower correlation value because fewer adolescents in the denominator as compared to adults or for reasons that need more study.
- Correlation between depression remission and depression response. The developer hypothesizes that clinics that have high response rates are also likely to have high remission rates for both adults and adolescents.
    - R-squared (adults) = 0.9051
    - R-squared (adolescents) = 0.7896
- Correlation between depression rates and rates of follow-up with a PHQ-9/9M. The developer hypothesizes that patients who receive regular screening are more likely to achieve remission for both adults and adolescents.
    - R-squared (adults) = 0.7967
    - R-squared (adolescents) = 0.7924
- Correlation between patients who achieve remission and patients who achieve response but not remission. The developer hypothesizes that clinics that have high response rates are also likely to have low response with no remission rates for both adults and adolescents.
    - R-squared (adults) = 0.3578
    - R-squared (adolescents) 0.2366
- Correlation between patients with depression outcome and diabetes outcome. The developer hypothesizes that there will be a weak but positive correlation between these two chronic conditions for adults only. R-squared (adults) = 0.1406

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

- This measure excludes patients who die, are a permanent resident of a nursing home or are enrolled in hospice are excluded from this measure, and patients who have a diagnosis of bipolar or personality disorder, schizophrenia or psychotic disorder, or pervasive developmental disorder.
- The developer lists bipolar diagnosis and active Schizophrenia, or Psychotic Disorder as required exclusions and the remaining exclusions as allowable. The developer states that because this is a

longitudinal measure the allowable exclusion may occur during the course of the measurement period.

- The developer reports that the exclusion rate for 140,099 patients was 3.45% in an analysis from 2020.

22. **Risk Adjustment**

22a. **Risk-adjustment method**

☐ None (only answer Question 20b and 20e) ☒ Statistical model ☐ Stratification

☐ Other method assessing risk factors (please specify)

22b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

22c. **Social risk adjustment:**

22c.1 Are social risk factors included in risk model? ☒ Yes ☐ No ☐ Not applicable

22c.2 Conceptual rationale for social risk factors included? ☐ Yes ☐ No

22c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

22d.**Risk adjustment summary:**

22d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

22d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☐ No

22d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

22d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☐ No

22d.5.Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

22e**. Assess the risk-adjustment approach**

- The measure is risk adjusted using a logistic regression model to create an indirect standardization risk adjustment (expected value). Performance is measured against the expected value for the given case mix of the clinic. Separate models were run for adults and adolescents.
- Risk variables included in the model include initial PHQ-9/PHQ-9M score, insurance product and patient neighborhood deprivation index (based on zip code). Deprivation index is new in 2021.
- The developer considered race, ethnicity, language and country of origin variables for the model. They did have an impact on the score, but the developer did not believe there was a conceptual basis for their inclusion and the potential for implicit bias. The social deprivation index was included as a proxy for the social determinants of health.
- Model discrimination statistics were not included; the developer provided the model estimates, but not a c-statistic or other model statistic.
- No concerns.

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

For cost/resource use measures, does this measure identify meaningful differences about cost and resource use between the measured entities?

- Variability of rates among medical groups around the statewide average was as follows:
  - Adults: 17.0% (range 0% to 37.2%), using 120,344 patients from 550 clinics
  - Adolescents: 14.5% (range 0% to 29.1%), using 11,658 patients from 118 clinics

- The developer reports that twelve month follow-up has the widest variation among medical groups, and that overall rates are low.
- The developer does not describe the statistical methods for identifying meaningful differences.
- The developer provided information in 2b.28 on the risk adjusted results. In the adult model, 85 of the 550 facilities performed above expectations, 106 performed below expectations and 359 performed as expected. In the adolescent model, 111 of 118 facilities performed as expected, while two facilities performed below expectations and five facilities performed above expectations

24. **Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.**
    - Not applicable.
25. **Please describe any concerns you have regarding missing data.**
    - Missing follow up data is included in the denominator and patients who are not re-assessed are treated as if they are not in remission.
    - The developer states that MN has made incremental improvements in rates of follow-up PHQ-9 at 12 months, from 17.0% in 2010 to 41.8% in 2019 for adults. Adolescents, a new population for this measure have a 2019 follow-up rate of 38.9%.
    - The developer states that missing data (follow-up PHQ-9 patient reported outcome assessment) is not an issue as those patients who are not re-assessed in follow-up remain in the denominator and are treated as if they are not in remission, but that low outcome rates are not solely attributed to lack of follow-up. A portion of patients are still experiencing symptoms of depression and are not in remission. A separate analysis for patients who were assessed with a follow-up PHQ-9 demonstrates that remission was at 24 percent while significant depression symptoms persisted for 49 percent of the patients (24% moderate, 15% major, and 10% severe).
        - There is a companion related measure that allows medical groups to understand their use of the PHQ-9/ PHQ-9M tool, NQF # 0712 Depression Utilization of PHQ-9M (also under maintenance review this cycle). This measure reports the rate of tool administration for patients with a diagnosis of depression or dysthymia seen during a four month
    - Missing follow up data is included in the denominator and patients who are not re-assessed are treated as if they are not in remission.

For cost/resource use measures ONLY:

If not cost/resource use measure, please skip to question 25.

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

    ☐ **Yes** ☐ **Somewhat** ☐ **No (If "Somewhat" or "No", please explain)**
27. **Describe any concerns of threats to validity related to attribution, the costing approach, carve outs, or truncation (approach to outliers):**
28. **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 accountable-entity level testing has been conducted)

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity 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 accountable-entity level and the patient/encounter level is required; if not conducted, should rate as INSUFFICIENT.)

29. **Briefly explain rationale for rating of OVERALL RATING OF VALIDITY and any concerns you may have with the developers' approach to demonstrating validity.**
    - All potential threats to validity are not empirically assessed – there is no demonstration of how the risk adjustment model fits the data (Box 1) à Rate as INSUFFICIENT

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

30. **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**

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

ADDITIONAL RECOMMENDATIONS

32. **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.**
    - No additional questions or concerns.

Criteria 1: Importance to Measure and Report

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]

Incorporation of adolescents into this measure results in additional guideline support and the addition of an patient reported outcome tool modified for adolescents (PHQ-9M)

[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:

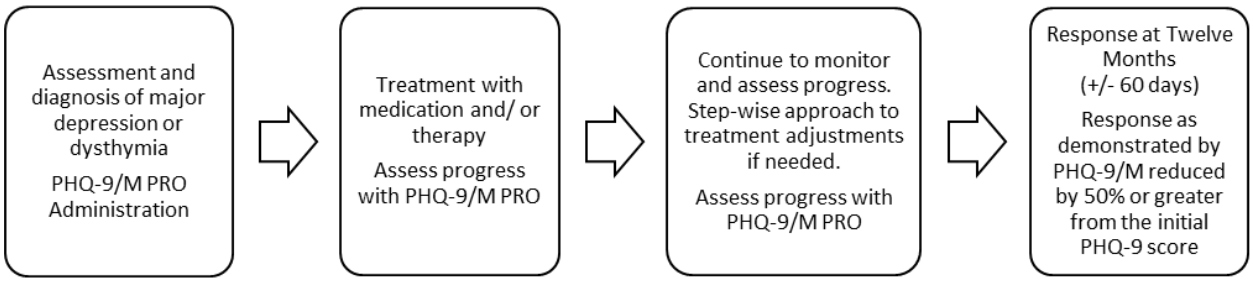
Evidence from the previous submission here.

1a. Evidence

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]



Health Care Process Steps to Achieve Desired Outcome for Depression

[Response Ends]

1a.02. Provide evidence that the target population values the measured outcome, process, or structure and finds it meaningful.

Describe how and from whom input was obtained.

[Response Begins]

Qualitative Study: Patients' perspectives on relevant treatment outcomes in depression treatment

Social functioning and interpersonal relationships The majority of patients mentioned goals related to social functioning (defined as an individual's ability to perform and fulfil normal social roles and interpersonal relationships as important goals of depression treatment. Normalization of social functioning was considered important (Table 2, quote 1). It included getting out of bed, continuing normal daily activities and functioning as before the depression. One patient stated that it was acceptable to use antidepressant medication, if necessary, for obtaining normalization of social functioning (Table 2, quote 2). Patients saw undertaking activities again with friends and family as a good indicator of social functioning. However, patients who had experienced multiple depressive episodes or patients who were diagnosed with chronic depression had a different view on functioning. They stressed that they needed to find new ways of functioning they would consider as satisfactory given circumstances, even though it would not quite be in the same way as before.

Themes	**Quotations for illustration**
Social functioning and interpersonal relationships	Quote 1: 'So the client's own picture of themselves [how the client themselves feels that they function], but also how those around them feel that they function. Because I think that's what's most important, if you can function more or less normally, like you used to.' (Participant 12, man, age 52) Quote 2: 'I was finally functioning without medication, and I thought that was fine. It is fine until another bump comes along and then you start all over again. If I ask myself now; I just want to be able to function again and, if necessary, with medication, like I did a few years ago. For me, that's my recovery.' (Participant 3, man, age 52)
Prevention of future recurrences	Quote 3: 'If you've been given um, enough things to hold on to to pull yourself up at times when you are sinking. Learning to recognize and know what you have to do about it. Identifying and tackling it.' (Participant 17, woman, age 25). Quote 4: 'Another way of dealing with it …, is to be able to relate success to your ability to deal with a setback yourself. Without having to go straight back into treatment or taking more pills, that when there are setbacks, a hard day, which in the past would have sent you straight into the abyss, now you have learned, first I have to do this and then I have to do that and watch out for this and so on…' (Participant 1, man, age 60)

Themes	Quotations for illustration
Acceptance of illness and managing the depression	Quote 5: 'During my first depressive episode, I really wanted things to be just like they were before. Although I did think that that would never happen, it was in fact my one sole wish. And, um, well, it's turned out be very different now from before, but better actually. But it was, it's been quite a process to accept things and to make adjustments.' (Participant 13, woman, age 41) Quote 6: 'I see recovery as learning to deal with your situation and to keep going. Because it will never make me better. And that has determined, and still determines, how I live my life and how I deal with my disabilities, what I do and what I don't do. Those are two aspects that the… um, come back every day. What do I do and what do I forget about? That's what, that's what it actually boils down to.' (Participant 16, woman, age 69)
Personal goals and societal expectations	Quote 7: 'That you go shopping, go to work and have a social life, and that this can be too much for people, or whether your goals is in fact that you can at least have a social life again, or just go to work, that can differ from one client to the next. But the outside world says, you're not really part of things again unless you're working, and that's what I'd really like to do.' (Participant 3, man, age 52) Quote 8: 'There is, for example, another goal that I have: in my contact with others I want to be less troubled by certain things, but that's not the same as not having any symptoms any more. And in my view, a practitioner often tends to look from that perspective, if things are x and y, then z is automatically the case, whereas it isn't always like that. Sometimes I can feel really good.' (Participant 2, woman, age 22) Quote 9: 'For almost everyone I can think of an example, with all the questionnaires [routine outcome monitoring questionnaires/symptom rating scales] that you have to fill in, that at some time they say, oh, you're doing a lot better, and that you definitely don't feel that yourself. So um, that's not the whole story.' (Participant 1, man, age 60)

Table 2 Quotes for each theme from the patient's perspective

Patients' and clinicians' perspectives on relevant treatment outcomes in depression: qualitative study Kaying Kan, Frederike Jörg, Erik Buskens, Robert A. Schoevers and Manna A. Alma. BJPsych Open (2020) 6, e44, 1–7. doi: 10.1192/bjo.2020.27

[Response Ends]

1a.03. Provide empirical data demonstrating the relationship between the outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

[Response Begins]

Institute for Clinical Systems Improvement (ICSI) Clinical Practice Guideline

Summary Table of Recommendations for Major Depressive Disorder and Persistent Depressive Disorder

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.

* Cell intentionally left empty

Institute for Clinical Systems Improvement Clinical Practice Treatment Guidelines

Establish Follow-Up Plan

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.

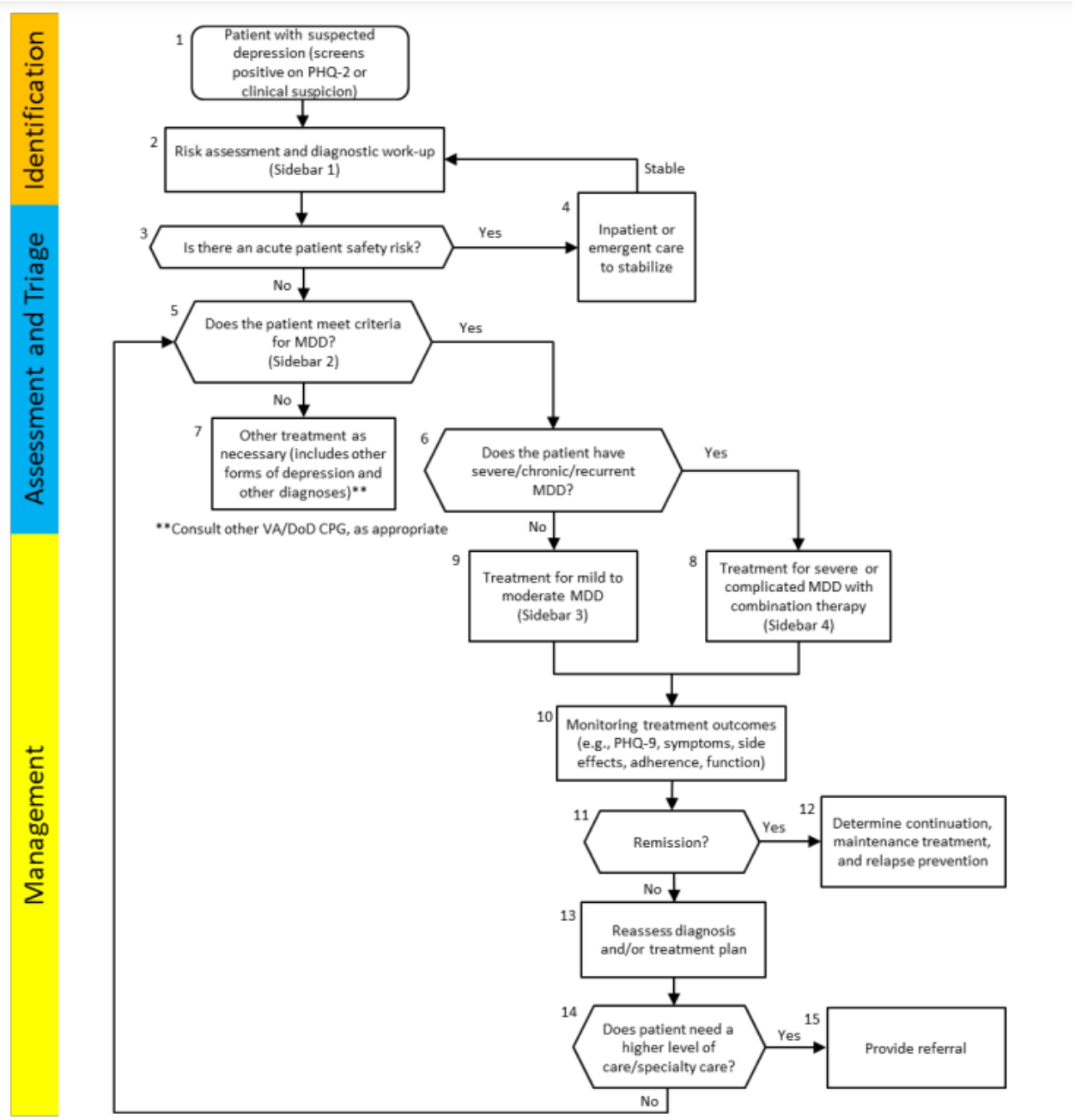
ICSI Institute for Clinical Systems Improvement Health Care Guideline Depression in Primary Care. Trangle M, Gursky J, Haight R, Hardwig J, Hinnenkamp T, Kessler D, Mack N, Myszkowski M. Institute for Clinical Systems Improvement. Adult Depression in Primary Care. Updated March 2016.

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

The image above depicts quality of evidence and strong recommendation for the importance of establishing a follow-up plan with the patient and maintaining contact with the patient as they continue treatment for depression symptoms.

ICSI Institute for Clinical Systems Improvement Health Care Guideline Depression in Primary Care. 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.

VA/DoD Major Depressive Disorder Clinical Practice Guideline



VA Department of Defense Clinical Practice Guidelines for Depression

Sidebar 3 Considerations in the Treatment of Mild/ Moderate MDD

For example:

a. Select monotherapy or combination therapy: pharmacotherapy/psychotherapy
b. Treatment for special populations (e.g., treatment of co-occurring conditions, pregnant patients, geriatric patients)
c. Patient preferences (treatment refusers)
d. Consider referral

Sidebar 4 Considerations in Treatment of Severe MDD

For example:

a. Recommend referral to specialty level of care
b. Select combination therapy: pharmacotherapy/psychotherapy
c. Treatment for special populations (e.g., treatment of co-occurring conditions, pregnant patients, geriatric patients)

d. Patient preferences (treatment refusers)

The image above depicts the treatment algorithm for depression from the Veteran's Administration Department of Defense which outlines the important components of screening, assessment, and recommended treatment based on severity of depression symptoms. Additionally, treatment recommendations are included for mild/ moderate depression and severe depression.

https://www.healthquality.va.gov/guidelines/MH/mdd/MDDCPGClinicianSummaryFINAL1.pdf

[Response Ends]

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

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

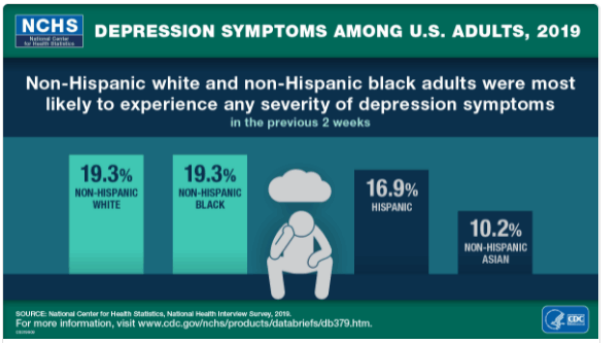
Explain how the measure will improve the quality of care, and list the benefits or improvements in quality envisioned by use of this measure.

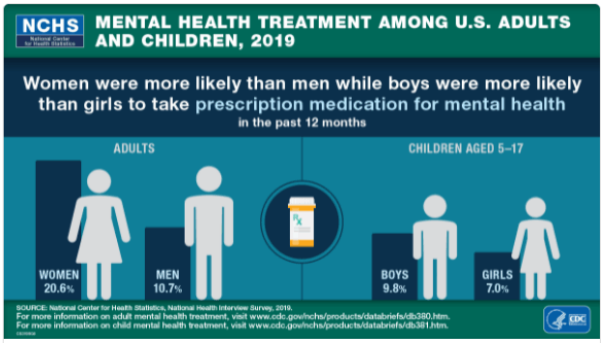
[Response Begins]

Adults:

Depression is a common and treatable mental disorder. The Centers for Disease Control and Prevention states that in 2019 (1)

- 2.8% of adults experienced severe symptoms of depression, 4.2% experienced moderate symptoms, and 11.5% experienced mild symptoms in the past 2 weeks.
- The percentage of adults who experienced any symptoms of depression was highest among those aged 18–29 (21.0%), followed by those aged 45–64 (18.4%) and 65 and over (18.4%), and lastly, by those aged 30–44 (16.8%).
- Women were more likely than men to experience mild, moderate, or severe symptoms of depression.
- Non-Hispanic Asian adults were least likely to experience mild, moderate, or severe symptoms of depression compared with Hispanic, non-Hispanic white, and non-Hispanic black adults.





Prevalence of Depression in Adults and Children; Centers for Disease Control 2019

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)

Adolescents:

- In 2019, 16% of the population ages 12–17 had at least one MDE during the past year, a higher prevalence than that reported in each year between 2004 (9%) and 2014 (11%).
- Among youth ages 12–17 in each year between 2004 and 2019, the prevalence of MDE was more than twice as high among females (ranging from 12% to 23%) as among males (ranging from 4% to 9%).
- The prevalence of MDE in 2019 was lowest among youth ages 12–13 (11%) compared with youth ages 14–15 (16%) and ages 16–17 (20%).
- Between 2004 and 2019, the prevalence of MDE increased for both genders among all three age groups (12–13, 14–15, and 16–17).
- The percentage of youth with MDE in the past year receiving treatment for depression increased between 2004 (40%) and 2019 (43%), but this increase was not statistically significant. Treatment was higher among females (46%) than among males (37%) in 2019. (5)

In 2015, 9.7% of adolescents in MN who were screened for depression or other mental health conditions, screened positively.

References

1. CDC. Symptoms of Depression Among Adults: United States, 2019 https://www.cdc.gov/nchs/data/databriefs/db379-H.pdf
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. CDC Children's National Indicators of Well-Being, 2021- Adolescent Depression https://www.childstats.gov/americaschildren/health4.asp

[Response Ends]

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

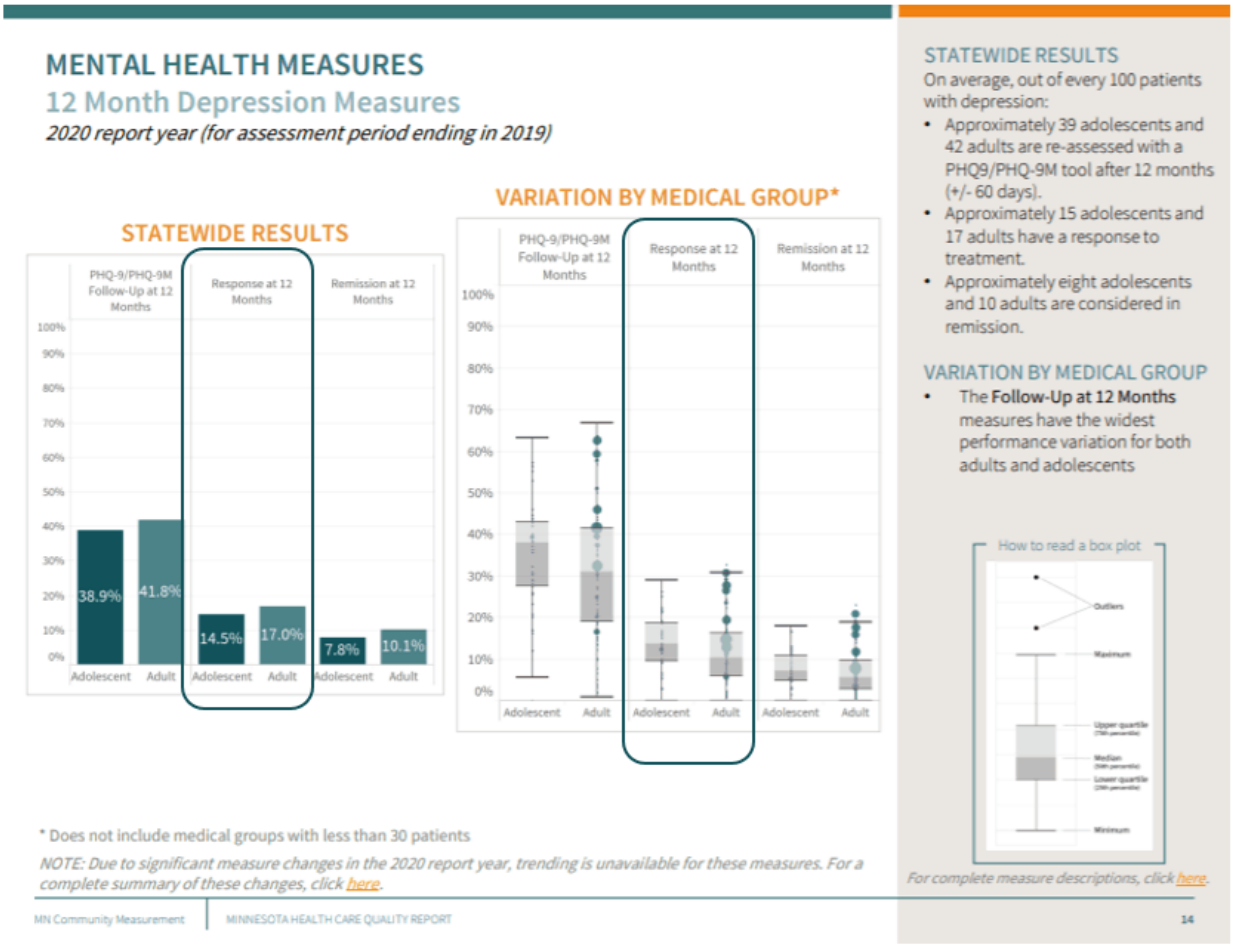
Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement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

[Response Begins]

Minnesota Statewide Reporting

Depression Response at 12 Months:

- Adults 17.0% (range 0% to 32.7%) 120,344 patients from 550 clinics
- Adolescents 14.5% (range 0% to 29.1%) 11,658 patients from 118 clinics



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

Above image depicts the statewide rates that demonstrate both opportunity for improvement (very low rates) and wide variability between clinic site results. Box plot diagram further displays the range and variability with several clinics achieving rates in the upper quartile box as well as several clinics in the lower quartile ranges.

Unable to provide trend data over the lifecycle of this measure due to significant redesign of the measure construct effective in the 2020 report year. However, a two year comparison is provided in an additional report for understanding the impact of COVID-19 on measure outcomes.

Summary of Depression Measure Changes

The following changes were implemented during the 2020 report year:

Change	Previous Report Year	Current Report Year
Age Criteria	18 years and older at time of encounter	12 years and older at time of encounter
Expansion of followup window	+/- 30 days • 6-month measures: 5 – 7 months • 12-month measures: 11 – 13 months	+/- 60 days • 6-month measures: 4 – 8 months • 12-month measures: 10 – 14 months
Acceptable PRO tool	PHQ-9 only	PHQ-9 or PHQ-9M (regardless of age)
Required Exclusions	• Bipolar disorder • Personality disorder	• Bipolar disorder • Schizophrenia/psychotic disorder
Allowable Exclusions	• Permanent nursing home resident • Hospice/palliative care • Death	• Permanent nursing home resident • Hospice/palliative care • Death • Personality disorder – emotionally labile • Pervasive developmental disorder

Change	Previous Report Year	Current Report Year
Behavioral health provider	Diagnosis of major depression or dysthymia must be in the primary position for encounters in a behavioral health setting.	No restrictions on major depression or dysthymia diagnosis positioning for behavioral health providers.
Allowable timing of PHQ-9 /PHQ-9M	PHQ-9 score at the time of encounter	PHQ-9/PHQ-9M score at time of encounter or up to seven days prior

https://mncm.org/reports/#community-reports

Issue Brief Depression Care in 2020 for Adults and Adolescents

KEY FINDINGS

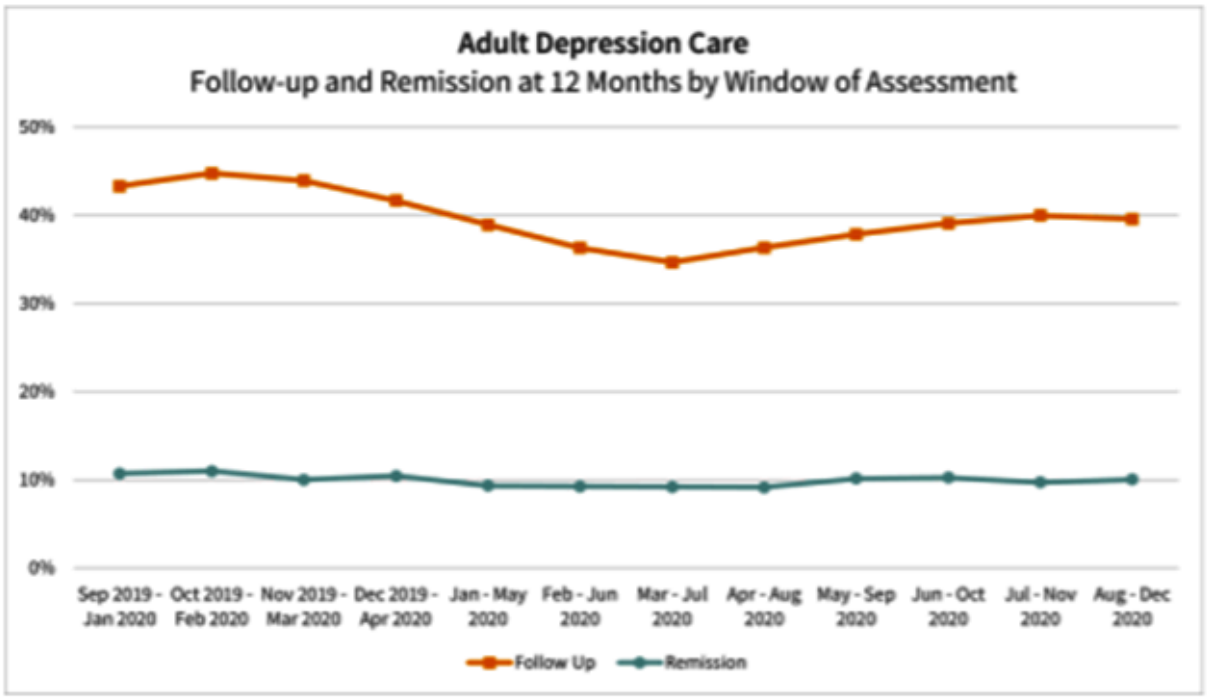
Adults

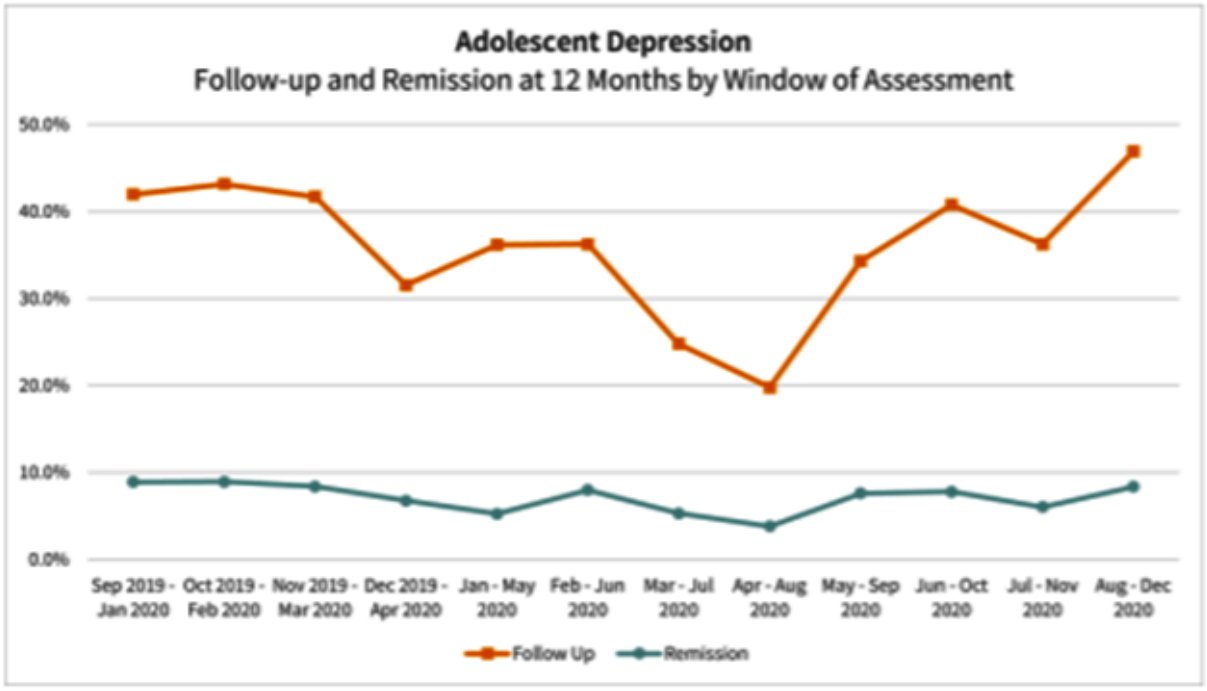
- Statewide, the PHQ-9/PHQ-9M Follow-up at 12 Months rate among adults decreased from 41.8% in 2019 to 39.6% in 2020. Additionally, the Remission at 12 Month rate among adults decreased from 10.1% in 2019 and 9.9% in 2020.
- In general, all demographic categories showed a decrease in both follow-up and remission rates between 2019 and 2020 for adults.
- The groups who experienced a significant worsening in their existing disparities were patients with the following demographic characteristics:
    - Follow-up at 12 Months: Asian, Native Hawaiian/Other Pacific Islander, on MHCP insurance and patients who are uninsured
    - Remission at 12 Months: Patients who are uninsured
    - Additionally, disparities worsened in some regions more than others.

Adolescents

- Statewide, the PHQ-9/PHQ-9M Follow-up at 12 Months rate among adolescents decreased from 38.9% in 2019 to 35.6% in 2020. Additionally, the Remission at 12 Month rate among adults decreased from 7.8% in 2019 and 7.0% in 2020.
- In general, all demographic categories showed a decrease in both follow-up and remission rates between 2019 and 2020 for adolescents.
- The groups who experienced a significant worsening in their existing disparities were patients with the following demographic characteristics:
    - Follow-up at 12 Months: Asian, males and on MHCP insurance
    - Remission at 12 Months: Asian
    - Additionally, disparities worsened in some regions more than others.

Since the design of the depression care measures tracks patients by the period of time in which they return the clinic for follow-up at 12 months, it is possible to see the impact at specific times during 2020.





Findings from Issue Brief Focus on Depression Measures- Impact of Pandemic COVID-19

The orange line/ square markers shows the percentage of patients that received a follow-up PHQ-9/PHQ-9M tool at 12 months. The light blue line/ round markers shows the percentage of patients who were considered in remission at 12 months (PHQ-9/PHQ-9M score less than 5). For each patient, the assessment window is 12 months (+/-60 days) after their index date, which creates some overlap in the time periods in the graph.

As expected, in 2020, the second quarter and the beginning of third quarter saw the lowest follow-up rates. This corresponds with the height of the COVID-19 disruptions to health care delivery. Interestingly, for adults, the rate of remission did not see the same decline. Despite the denominator for the adolescent population being smaller and thus having more volatile changes in rates, the population followed a similar pattern as the adult population.

https://mncmsecure.org/website/Reports/Spotlight%20Reports/2020%20MY%20Issue%20Brief%20-%20Depression.pdf

[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]

N/A

[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.

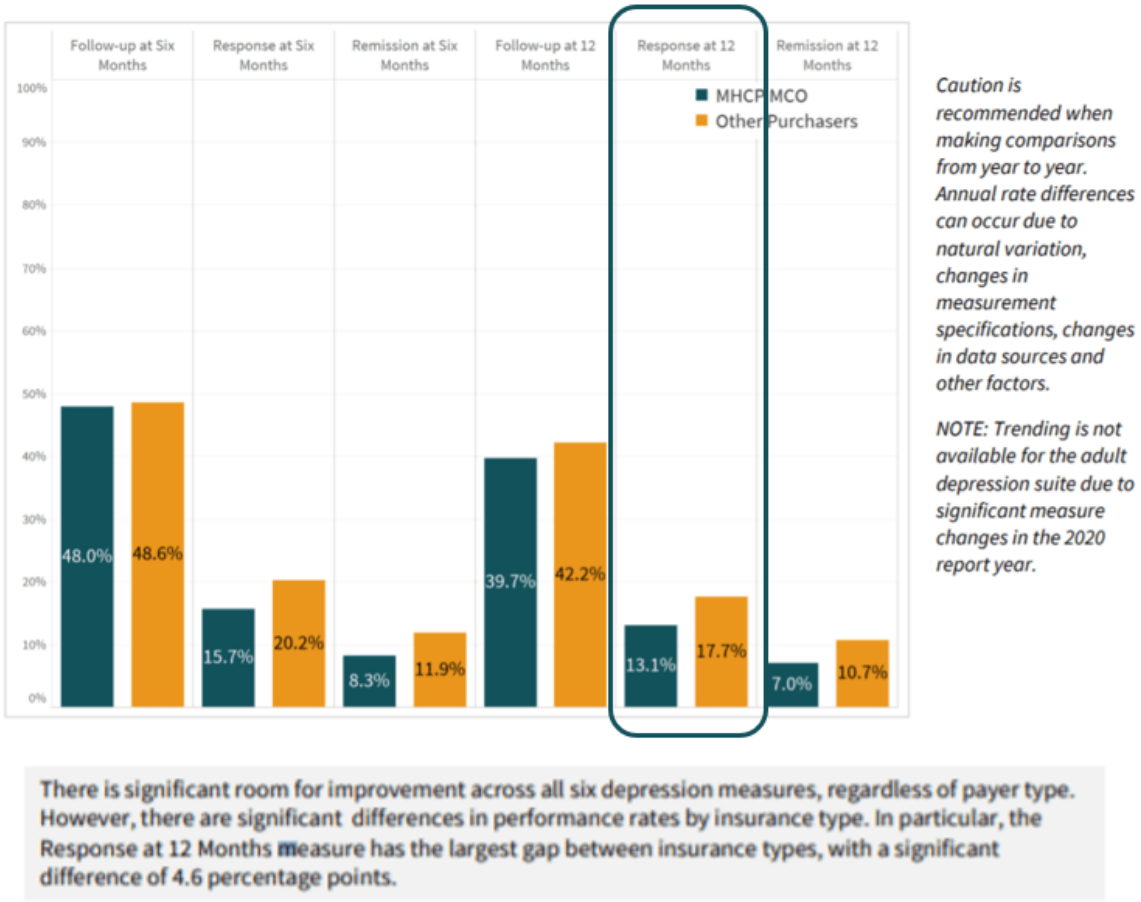
[Response Begins]

Minnesota Health Care Disparities Reports

Displays by MHCP MN Health Care Programs and Race

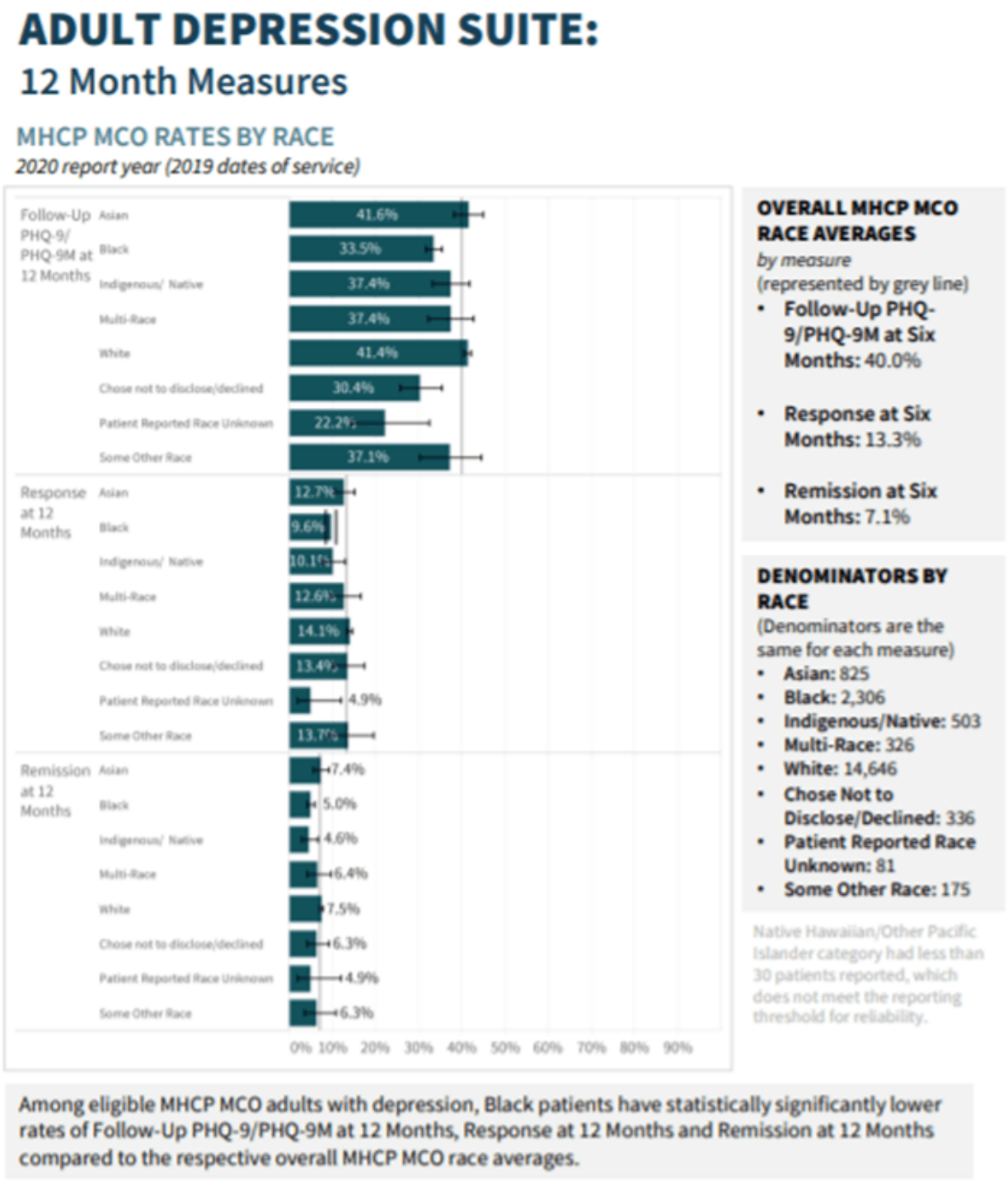
https://mncmsecure.org/website/Reports/Community%20Reports/Disparities%20by%20Insurance%20Type/2020%20RY%20Disparities%20by%20Insurance%20Type.pdf

https://mncmsecure.org/website/Reports/Community%20Reports/Disparities%20by%20RELC/2020%20Disparities%20by%20RELC%20Chartbook%20-%20FINAL.pdf



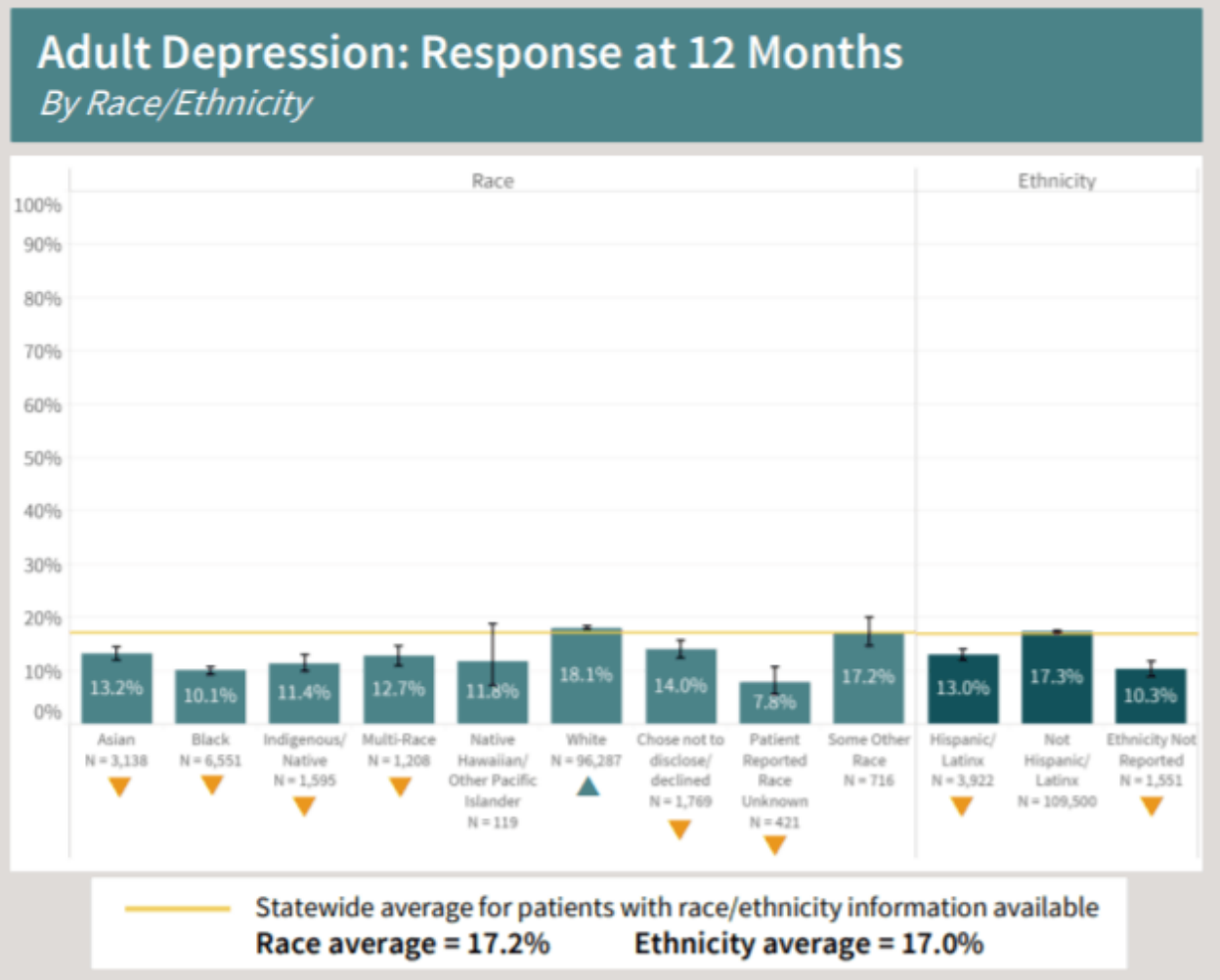
Minnesota Health Care Disparities Reports; Displays by MHCP MN Health Care Programs and Race

There is significant room for improvement across all six depression measures, regardless of payer type. However, there are significant differences in performance rates by insurance type. In particular, the Response at 12 Months measure has the largest gap between insurance types, with a significant difference of 4.6 percentage points.



Minnesota Health Care Disparities Reports; Displays by MHCP MN Health Care Programs and Race

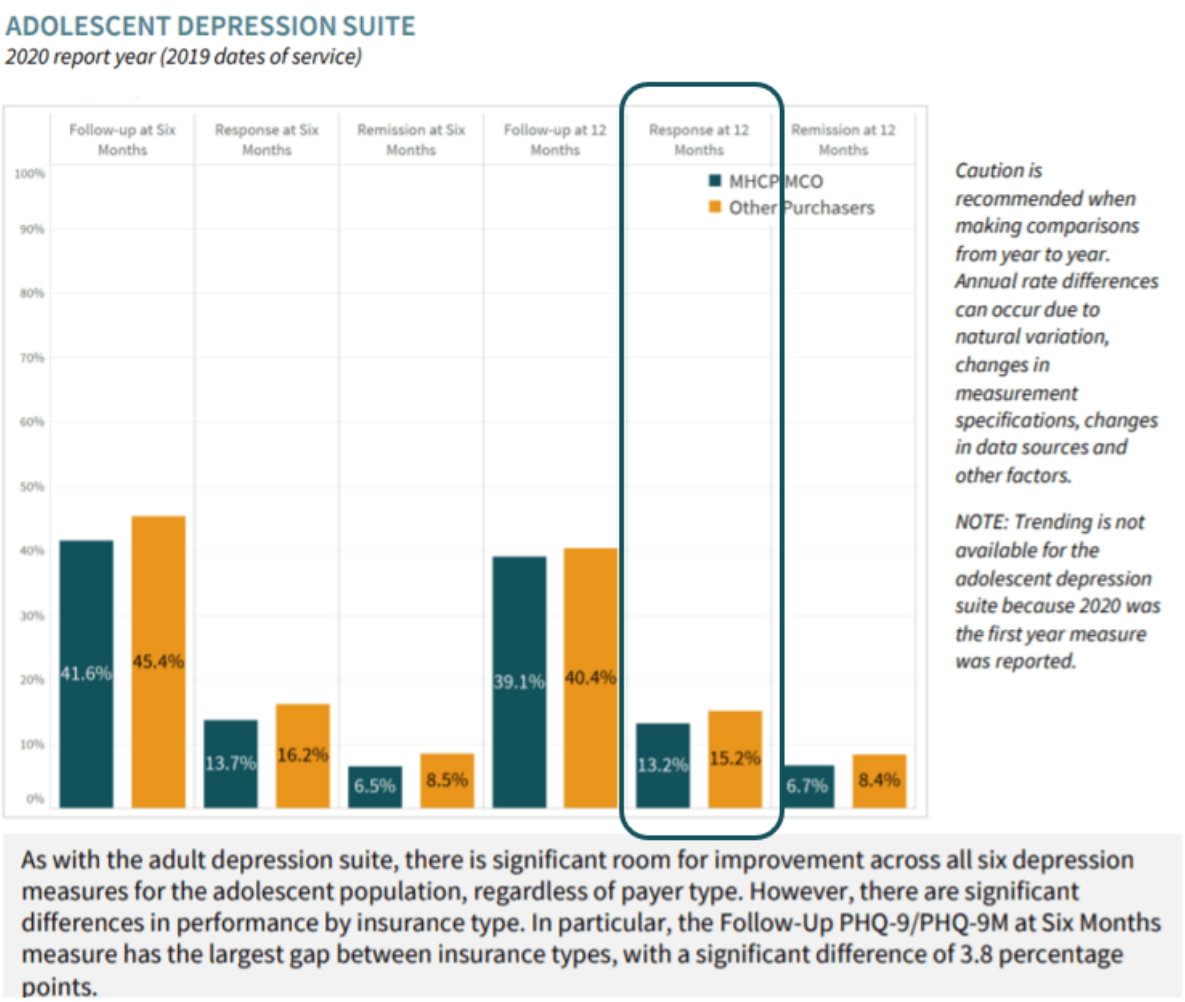
Among eligible MHCP MCO adults with depression, Black patients have statistically significantly lower rates of Follow-Up PHQ-9/PHQ-9M at 12 Months, Response at 12 Months and Remission at 12 Months compared to the respective overall MHCP MCO race averages.



Minnesota Health Care Disparities Reports; Displays by MHCP MN Health Care Programs and Race

The above bar graph depicts the rates of depression response at 12 months for adults by race as compared to the statewide average. A gold downward pointing triangle indicates that rates are significantly lower than average (Asian, Black, Indigenous/Native, Multi-Race and Hispanic/Latinx) as compared to the statewide average. A green colored upward pointing triangle reflects a rate significantly higher than average (White).

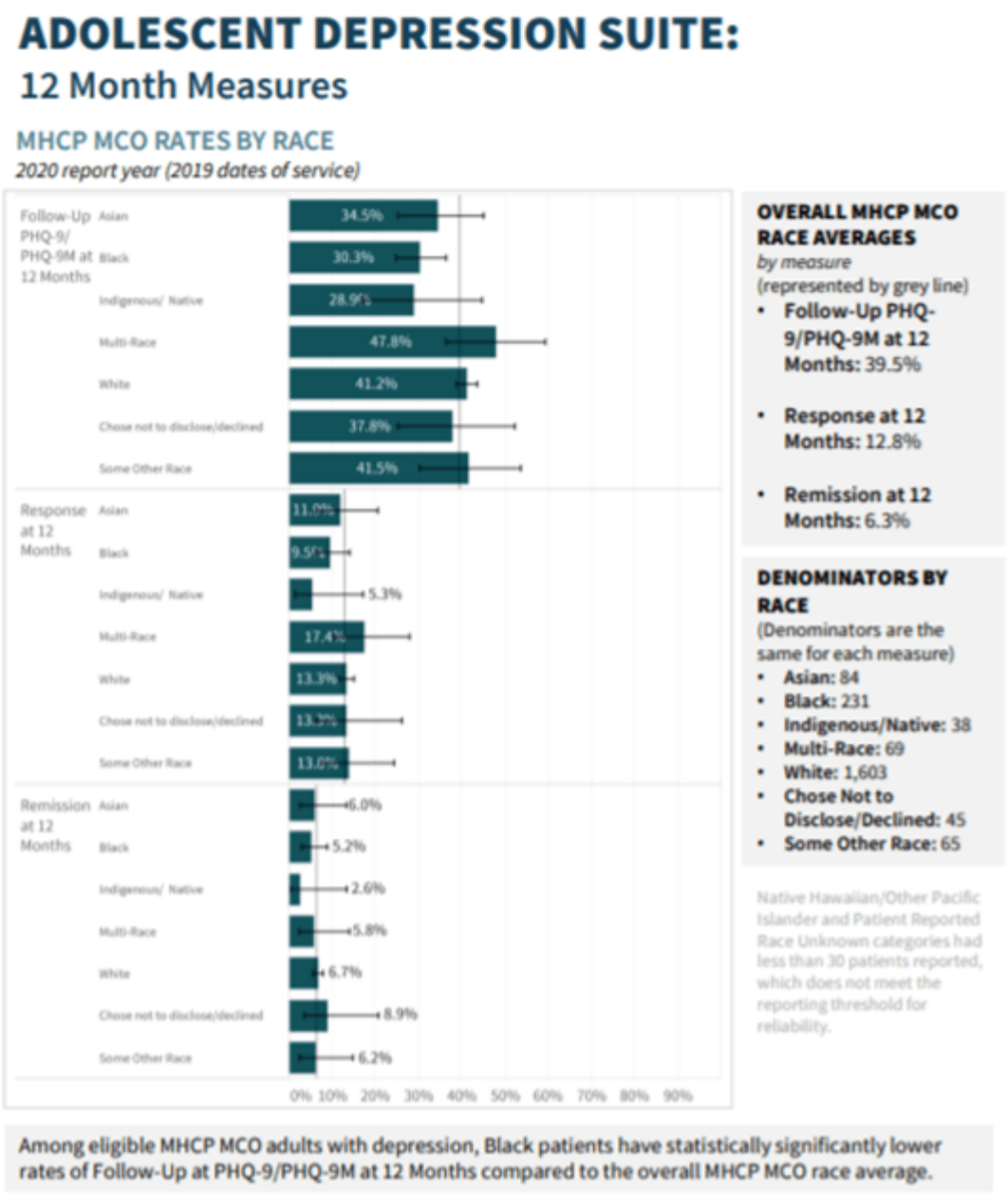
Patients who are Asian, Black, Indigenous/Native, Multi-Race or Hispanic/Latinx are among those who have significantly lower rates of depression follow-up, response and remission at 12 months compared to the race/ethnicity averages.



Minnesota Health Care Disparities Reports; Displays by MHCP MN Health Care Programs and Race

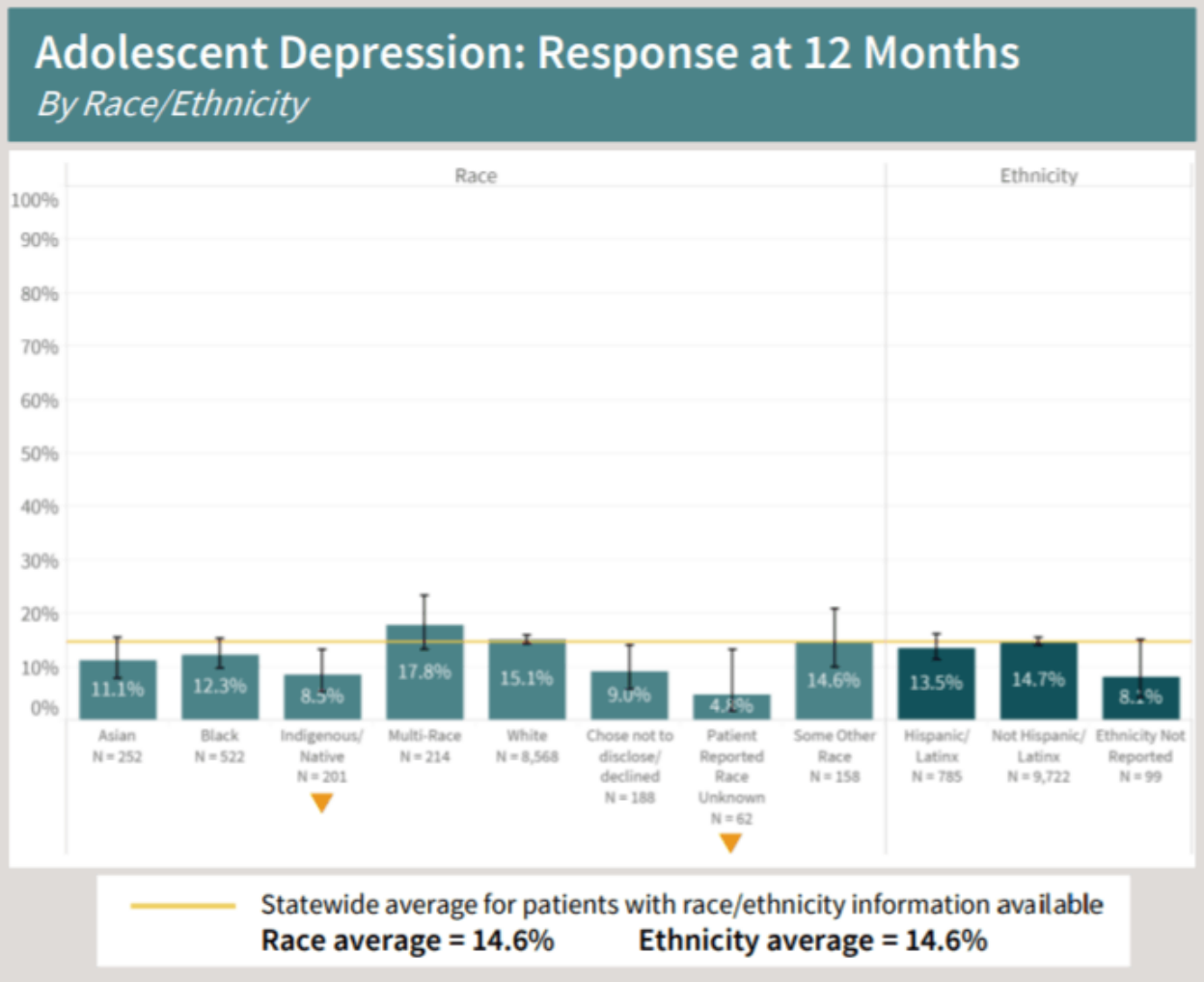
The above graph demonstrates a 3.0 percentage point difference (decrease) between patients with MHCP insurance versus all other payers.

As with the adult depression suite, there is significant room for improvement across all six depression measures for the adolescent population, regardless of payer type. However, there are significant differences in performance by insurance type. In particular, the Follow-Up PHQ-9/PHQ-9M at Six Months measure has the largest gap between insurance types, with a significant difference of 3.8 percentage points.



Minnesota Health Care Disparities Reports; Displays by MHCP MN Health Care Programs and Race

Among eligible MHCP MCO adolescents with depression, Black patients have statistically significantly lower rates of Follow-Up PHQ-9/PHQ-9M at Twelve Months and Response at Twelve Months compared to the respective overall MHCP MCO race averages.



Minnesota Health Care Disparities Reports; Displays by MHCP MN Health Care Programs and Race, Adolescents

The above bar graph depicts the rates of depression response at 12 months for adolescents by race as compared to the statewide average. A gold downward pointing triangle indicates that rates are significantly lower than average for Indigenous/Native and Patient Reported Race Unknown) as compared to the statewide average. A green colored upward pointing triangle reflects a rate significantly higher than average.

Although the measure does not demonstrate a high, topped out performance rate and demonstrates continued variability and opportunity for improvement, stratification by race/ ethnicity, and insurance indicate further opportunities for improvement.

[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]

N/A

[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
- expanding the assessment window to +/- 60 days
- modified exclusion value set for personality disorder
- added exclusions for schizophrenia and pervasive developmental disorder
- removed the requirement that the depression diagnosis be in the primary position for behavioral specialty

[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; widen the follow-up assessment window; 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. Widen the follow-up assessment window to +/- 60 days for all populations and all response and remission measures

* Six-month measure's assessment window expands from 5 to 7 months to 4 to 8 months

* Twelve-month measure's 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. 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.

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: 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.

5. 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.

Please refer to the data dictionary (sp.11) for the summary of redesign activities and changes to value sets or https://helpdesk.mncm.org/helpdesk/KB/View/22742768--depression-changes-and-rationale

[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 Response at Twelve Months- Progress Towards Remission

[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 major depression or dysthymia who are progressing towards remission by achieving a response (PHQ-9 or PHQ-9M score reduced by 50% or greater) twelve months (+/- 60 days) after an index visit.

[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: Change

[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

https://helpdesk.mncm.org/helpdesk/KB/View/20945873-risk-adjustment-how-is-the-expected-rate-calculated

[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: 1885_MNCM Depression Care VS Specs Definitions w Redesign 6-9-2021.xlsx

For the question below: state the outcome being measured. Calculation of the risk-adjusted outcome should be described in sp.22.

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]

The number of patients in the denominator who achieved a response as demonstrated by a PHQ-9 or PHQ-9M score reduced by 50% or greater twelve months (+/- 60 days) after an index visit.

[Response Ends]

For the question below: describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in sp.22.

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]

This PROM-PM outcome measure is longitudinal, seeking to measure improvement of depression symptoms with a PHQ-9 or PHQ-9M result reduced by 50% or greater (response) within twelve months (+/- 60 days) for the patients with an index event(depression and elevated PHQ-9 or PHQ-9M).

The numerator is defined as patients with a twelve-month (+/- 60 days) PHQ-9 or PHQ-9M score reduced by 50% or greater.

The numerator rate is calculated as follows:

pts with major depression or dysthymia with a PHQ-9 or PHQ-9M score reduced by 50% or greater at 12 months(+/- 60 days)/

pts with major depression or dysthymia with index contact PHQ-9 > 9

Patients who do not have a twelve month +/- 60 day PHQ-9 or PHQ-9M score obtained remain in the denominator and are counted as not having a response to treatment. Not having a PHQ-9 or PHQ-9M score within the 120 day window is considered a numerator miss.

Time period for data collection: there is a set index period for this measure, typically patients who have an index visit within a calendar period (e.g. index dates between 11/1/2017 and 10/31/2018) and then allowing enough time to pass to accommodate the timeframe for assessment. (e.g. for response at twelve months +/- 60 days with index dates of service ending 10/31/2018, the assessment period for twelve month remission and response [to also capture 12 month remission and response rates] would go through 12/30/2019). Technically, the six- and twelve-month remission and response measures are collected together in the MN program, and the index assessment period is fourteen months in duration.

Denominator identification period (index) 11/1/2017 to 10/31/2018

Measure assessment period through 12/30/2019; reported in 2020.

[Response Ends]

For the question below: state the target population for the outcome. Calculation of the risk-adjusted outcome should be described in sp.22.

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 major depression or dysthymia and an initial (index) PHQ-9 or PHQ-9M score greater than nine.

[Response Ends]

For the question below: describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in sp.22.

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 major depression or dysthymia and an initial (index) PHQ-9 or PHQ-9M score greater than nine, is identified as follows:

Patients age 12 and older at the time of the index visit

AND Index visit

An index visit occurs when ALL of the following criteria are met during a face-to-face visit or contact with an eligible provider:

- a PHQ-9 or PHQ-9M result greater than nine
- an active diagnosis of Major Depression or Dysthymia (Major Depression or Dysthymia Value Set)
- the patient is NOT in a prior index period

An index period begins with an index visit and is 14 months in duration.

Denominator is stratified by age range for adolescents (12 to 17 years of age) and adults (18 years of age and older).

Patients who do not have a twelve month +/- 60 day follow-up PHQ-9 or PHQ-9M score obtained remain in the denominator for this measure.

Please refer to the attached data dictionary for an inclusive list of all ICD-9/ ICD-10 codes and data element definitions.

[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 their measure assessment period
- Patient had an active diagnosis of Schizophrenia or Psychotic Disorder (*Schizophrenia Psychotic Disorder* Value Set) any time prior to the end of their measure assessment period

Allowable exclusions:

- Patient had an active diagnosis of Personality Disorder – Emotionally Labile Conditions (*Personality Disorder – Emotionally Labile* Value Set) any time prior to the end of their measurement assessment period
- Patient had an active diagnosis of Pervasive Developmental Disorder (*Pervasive Disorder* Value Set) any time prior to the end of the measurement assessment period
- Patient was a permanent nursing home resident at any time during the denominator identification period or measure assessment period
- Patient was in hospice or receiving palliative care (*Palliative Care* Value Set) at any time during the denominator identification or measure assessment period
- Patient died prior to the end of their measurement assessment 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]

Statistical risk model

[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]

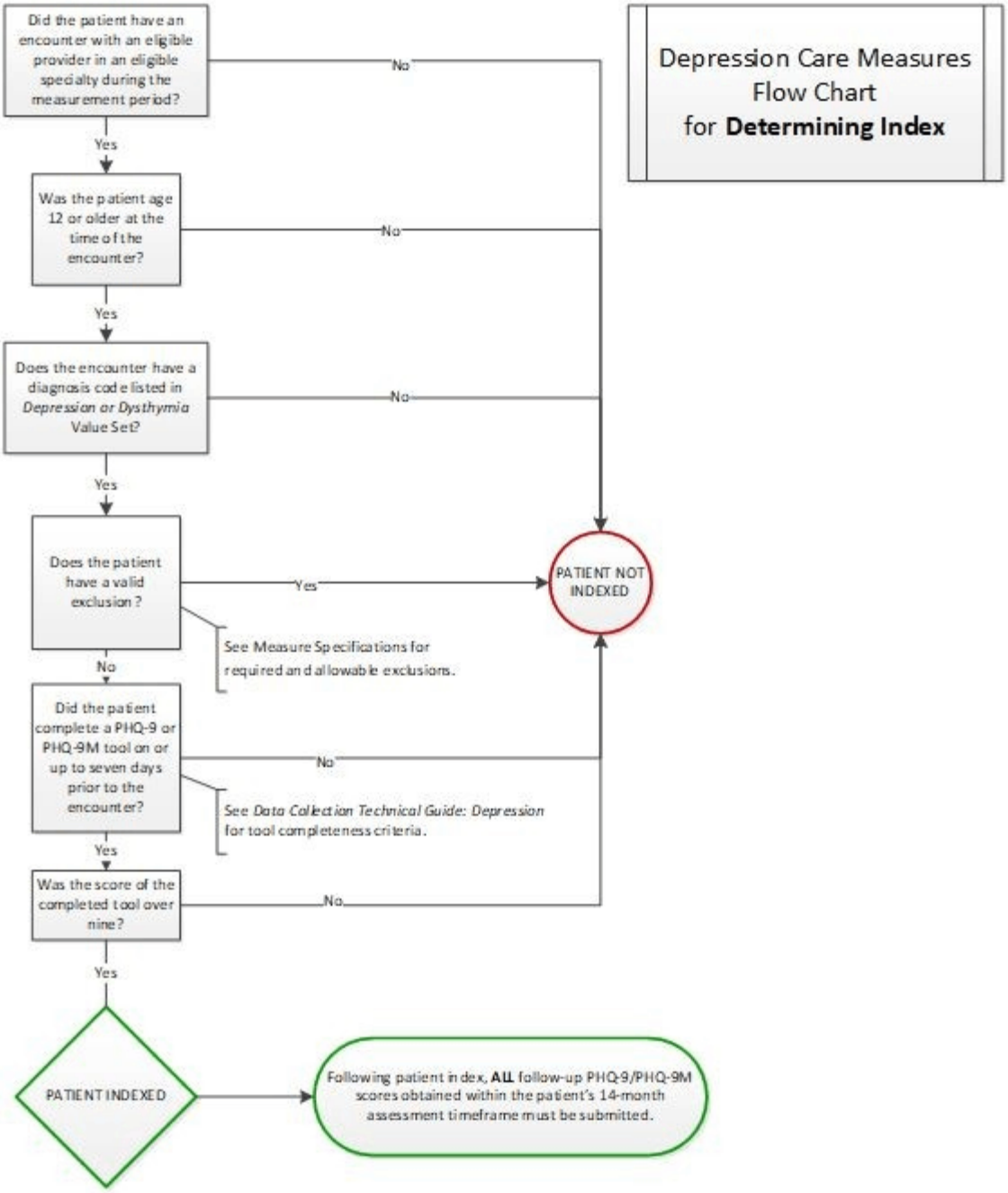
Better quality = Higher score

[Response Ends]

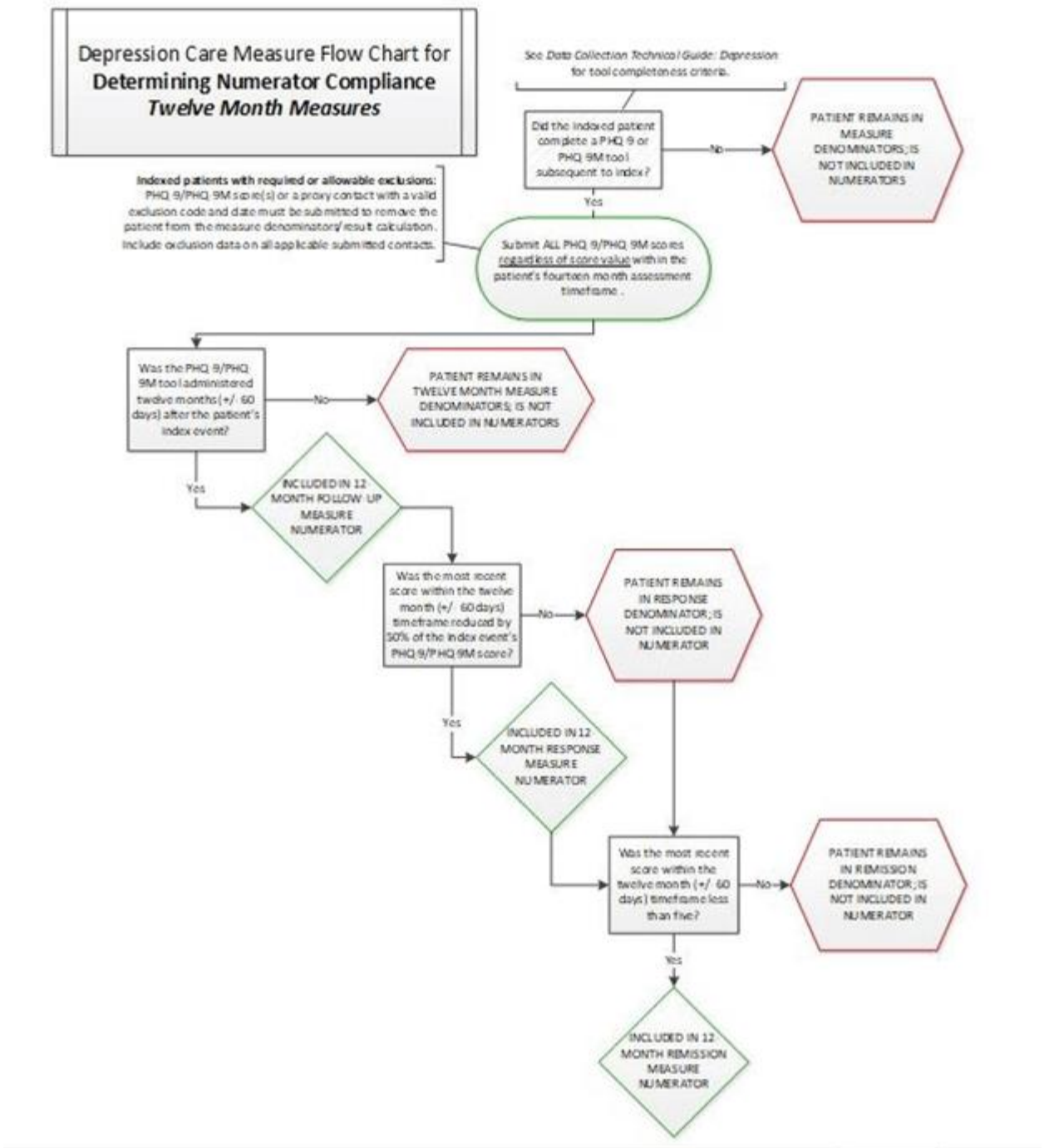
sp.22. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

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]



Measure Calculation Algorithms; Determining Depression Index and Calculation of Numerator



Measure Calculation Algorithms; Determining Depression Index and Calculation of Numerator

This measure is calculated by submitting a visit level file for the eligible patients. Each record in the file represents a contact with the patient and PHQ-9 or PHQ-9M score associated with this contact. Data files are submitted to a HIPAA secure data portal. Programming within the data portal determines the starting point (index visit) and then calculates based on dates if a twelve month +/- 60 days PHQ-9 or PHQ-9M was obtained and the resulting score.

Calculation logic:

Is patient eligible for inclusion with diagnosis codes (Major Depression or Dysthymia Value Set)

and PHQ-9 or PHQ-9M > 9?

If yes, mark the visit as index (anchor) and include this patient in the denominator.

Does patient have a PHQ-9 or PHQ-9M score completed with a contact date that is twelve months +/- 60 days from the index date?

If yes, include this score to calculate rate. Programming logic includes the most recent score within the +/- 60 day window.

If no, patient is included in the denominator only. Not having a PHQ-9 or PHQ-9M score within the 120 day window is considered a numerator miss.

If the patient does have a twelve month +/- 60 day PHQ-9 or PHQ-9M and the score is it reduced by 50% or more from the index PHQ-9 or PHQ-9M score? [For example, a patient with an index PHQ-9/PHQ-9M score of 21 then at twelve months +/- 60 days has a most recent follow-up score of 9 would be considered a response and in the numerator]

If twelve month +/- 60 day PHQ-9 or PHQ-9M is reduced by 50% or greater; is considered a numerator case for rate calculation.

Reporting of this measure is currently at the clinic and medical group level.

Risk adjustment methodology uses individual patient level variables (age, insurance product depression severity level and zip code based deprivation index) to adjust for these variables at the clinic site and medical group practice level. Age is a continuous variable. Insurance product is Commercial, Medicare, Minnesota Health Care Plans (MHCP) and Cash or Uninsured patients. Depression severity level is based on the index PHQ-9 or PHQ-9M score, Moderate (PHQ9 below 15), Moderately Severe (PHQ9 15 to 19), Severe (PHQ9 over 19). The risk adjustment employs an actual to expected methodology where the actual measure result remains unaltered, instead a risk adjusted comparison is created based on same proportions of the risk factors that the clinic has. Our MNHealthscores website displays both the actual and expected rates in the detailed view.

[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: 1885_PHQ9.pdf

Attachment: 1885_PHQ-9-Modified-For-Teens-64711 GLAD-PC.doc

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 for full population reporting the EMR as the data source. Not amenable to sampling because 1) each patient's starting point for measurement is different, depending on the date of elevated PHQ-9 and 2) the longitudinal nature of the measure tracking improvement over time.

[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 2008Vol. 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 am EMR. A CSV file is created by each medical group and uploaded to a password protected, HIPAA secure data portal which performs rate calculation. Selected Patient Reported Data, not because it is necessarily a separate data source, but because this measure is based on a patient reported outcome tool, a PRO-PM measure. Frequently this PRO tool, the PHQ-9, is housed within a clinic's EMR, or in paper charts is a part of the patient's medical record.

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]

Yes

[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]

Yes - Additional risk adjustment analysis is 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. Reliability

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. MNCM receives patient level data via a HIPAA secure data portal, so each year data is available for reliability, validity and risk adjustment variable testing on a large population. For this measure, due to its longitudinal 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

Denominator identification period (index) 11/1/2017 to 10/31/2018

Measure assessment period through 12/30/2019; reported in 2020

Measure Assessment Period: For each patient, the measure assessment period begins with an index event and is 14 months (12 months +/- 60 days) in length. The assessment period is held constant to assess the same denominator of eligible patients for outcomes of remission and response at both six and twelve months.

[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

Clinician: Individual

[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 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. Over 115 medical groups representing 788 clinics were included in the testing of this measure, representing 118,132 adults and 7,237 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

118,132 adult patients and 7,237 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 ≥ 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]

Social risk factors available and analyzed for this measure include age, race, ethnicity, primary language, country of origin and zip code-based deprivation index.

[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 ≥ 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).

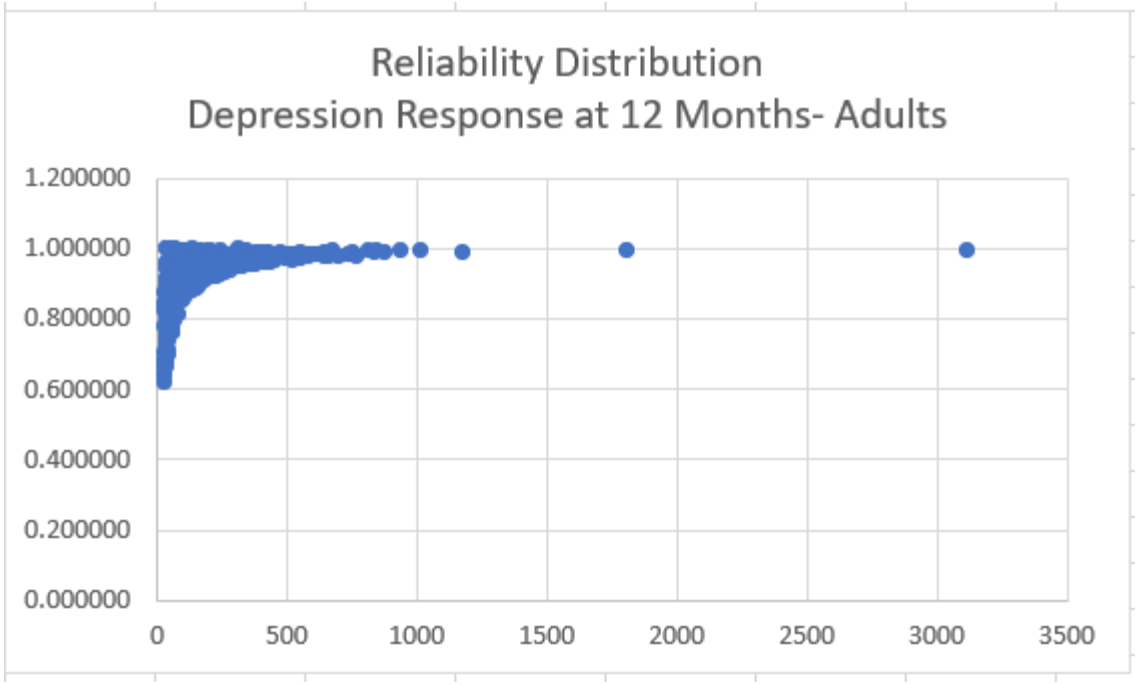
[Response Begins]

2021 Submission

Adults age 18 and older

550 clinics, 118,132 patients

Average Reliability score: 0.921229

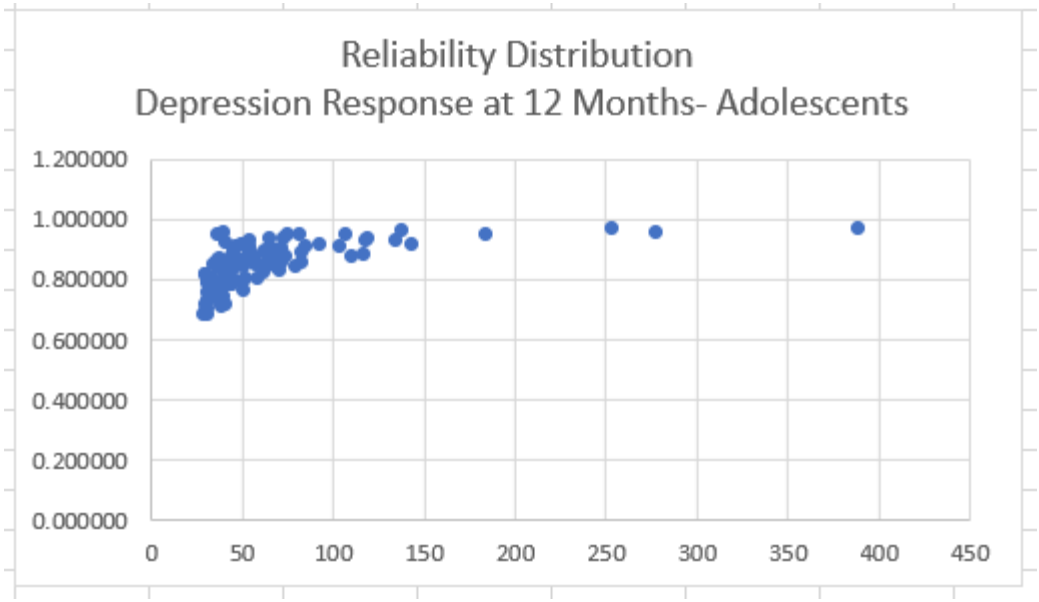


2020 Beta-binomial Reliability Performance Score- Adults 0.92129 (# of clinics = 550, number of patients = 118, 132)

Adolescents age 12 to 17

118 clinics, 7,237 patients

Average Reliability Score: 0.836776



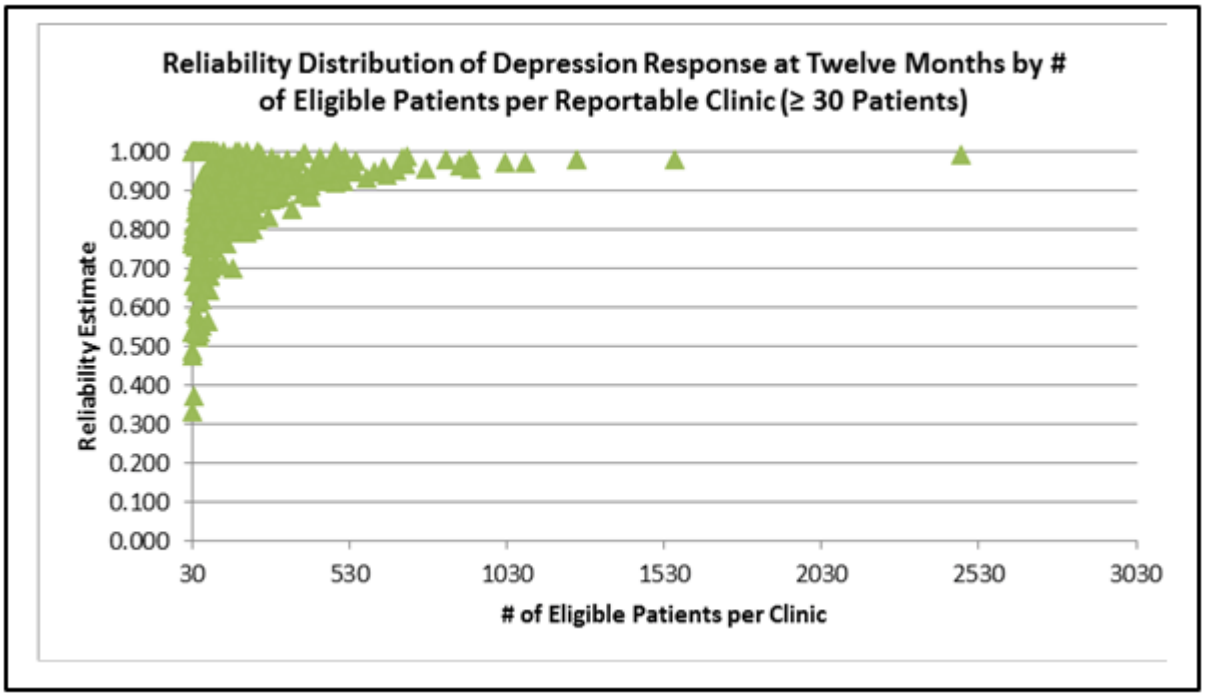
2020 Beta-binomial Reliability Performance Score- Adolescents 0.836776 (# of clinics = 118, number of patients = 7.237)

2013 Submission

Reliability = 0.881

Reportable clinics (≥ 30 patients)

- α = 1.6217
- β = 18.5926
- σ^2 (provider to provider variance) = 0.00348
- average reliability = 0.881



2013 Original Beta-binomial Reliability Performance Score- Adults 0.881

[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

- **Relia**bility score = 0.921229 (Adult) and 0.836776 (Adolescents)

For clinics reporting measure results for adults (550 clinics and 118,132 patients), the reliability performance score was calculated at 0.915806. 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 0.91, there is the ability to distinguish higher performing clinics from lower performing clinics.

It is noted that the reliability performance score increased with the changes made to the measure during the redesign process (enhanced exclusions and widening the assessment window to +/- 60 days.)

Although there are fewer clinics reporting measure results for adolescents (118) and fewer adolescents (7,327) as compared to the adult population, the reliability performance score is still quite high at 0.836776. 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. Validity

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
- Attestation of inclusion of patients both with newly diagnosed depression and those with existing depression and elevated PHQ-9
- 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.

Validation audits verify that the clinical data submitted for the numerator component of the measure matched the data in the patient record. Other data elements are also audited to verify the patient was included in the denominator correctly (e.g., diagnosis of depression).

Validity Performance Score: Correlation was performed against several different measures. Interpretation of correlation statistics is as follows:

- 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).
- High degree: If the coefficient value lies between ± 0.50 and ± 1, then there is said to be a strong correlation.

Hypotheses tested included:

1. The correlation between two similar depression outcome measures; depression remission (PHQ-9 < 5) and depression response (PHQ-9 $\underline{>}$ 50 percent improved from index initial PHQ-9 score). The hypothesis is that clinics who do well achieving the response outcome will also do well at achieving remission. Clinically, patients with depression who have a response to treatment don't always reach remission, but the clinic-level measure rates should show some correlation.
2. The correlation between depression outcome rates and the rates of follow-up with a PHQ-9/ PHQ-9M. Patients who have regular follow-up PHQ-9 assessments with their providers represent ongoing evaluation of the patient's treatment plan and are more likely to achieve remission (PHQ-9 or PHQ-9M < 5) or a response to treatment (PHQ-9 or PHQ-9M is equal to or greater than 50% improved from index PHQ-9).
3. The correlation between patients who achieve remission and those who achieve response but not remission. This is an enhancement to the hypothesis stated in #1 in that it separates the measure rates into two distinct populations.
4. For the adult population, the correlation between depression outcome measures and another chronic condition measure for a diabetes composite measure. The hypothesis is expected to be somewhat weak because the conditions of depression (chronic-episodic) and diabetes (chronic) reflect different clinical course of care, different outcomes, and a different measure construct. However, there may be some correlation.

[Response Ends]

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

Examples may include correlations or t-test results.

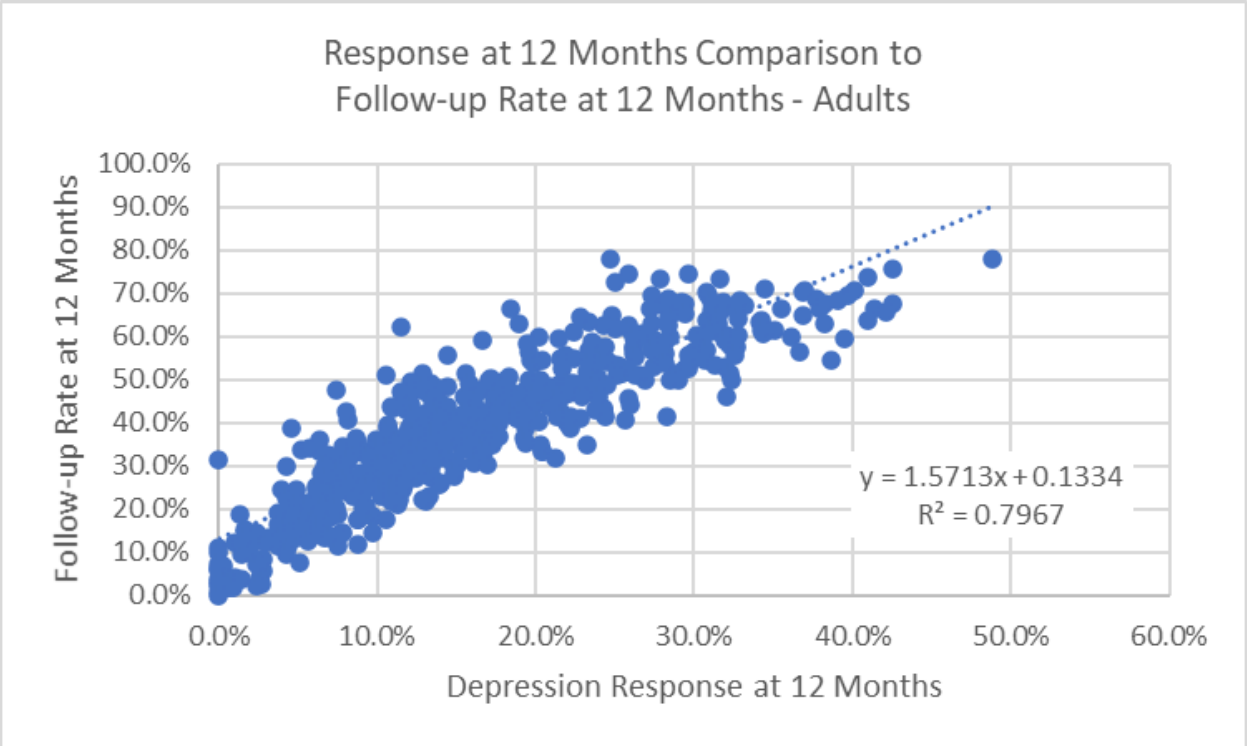
[Response Begins]

2021 Submission

Validity Performance Scores (Correlation)- Adults

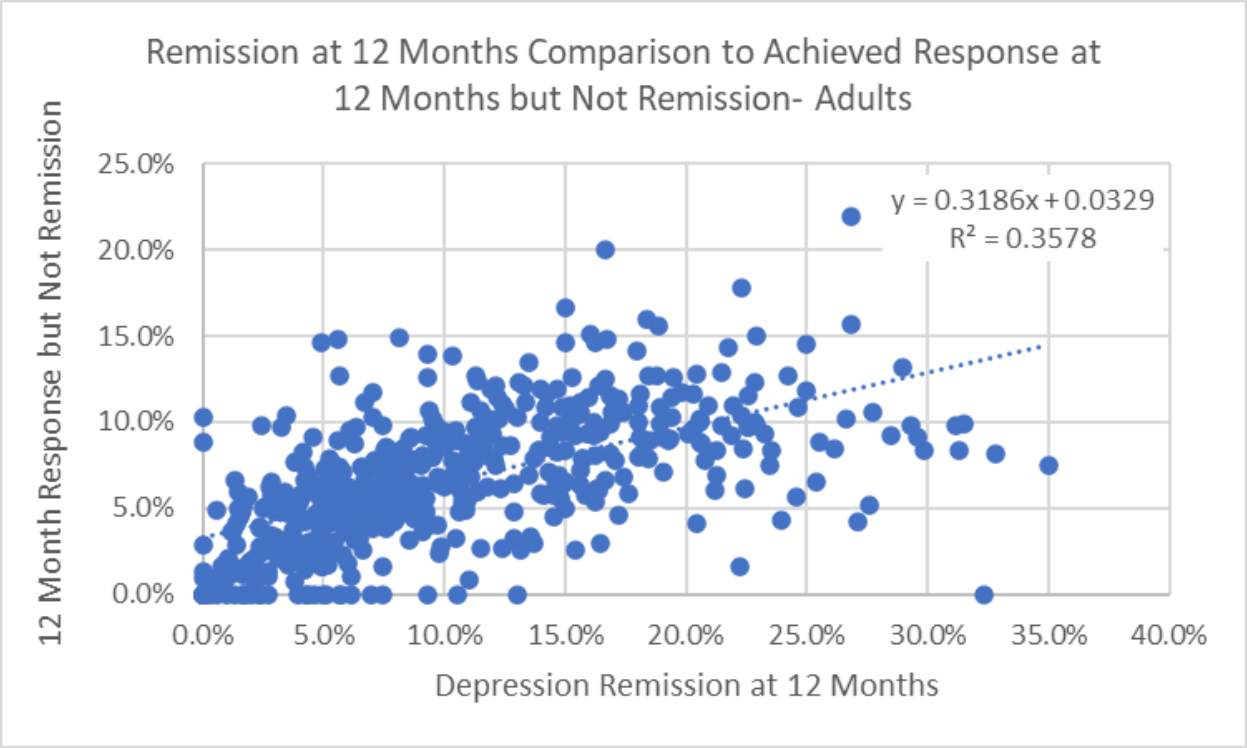
Hypothesis	Description	r-squared
#1	Correlation between depression remission (PHQ-9 < 5) and depression response (PHQ-9 ≥ 50 percent improved from index initial PHQ-9 score)	0.9051
#2	Correlation between depression outcome rates and the rates of follow-up	0.7967
#3	Correlation between patients who achieve remission and those who achieve response but not remission	0.3578
#4	Correlation between depression outcome and a diabetes composite measure	0.1406

Display of Hypothesis #2 Correlation between Depression Response at 12 Months and Follow-up Rate at 12 Months



Correlation between Depression Response at 12 Months and Rate of Follow-up at 12 Months; 550 clinics and 118,132 patients R-Squared value of 0.7967

Display of Hypothesis #3 Correlation between Depression Response at 12 Months and Patients who Achieve Response but Not Remission at 12 Months

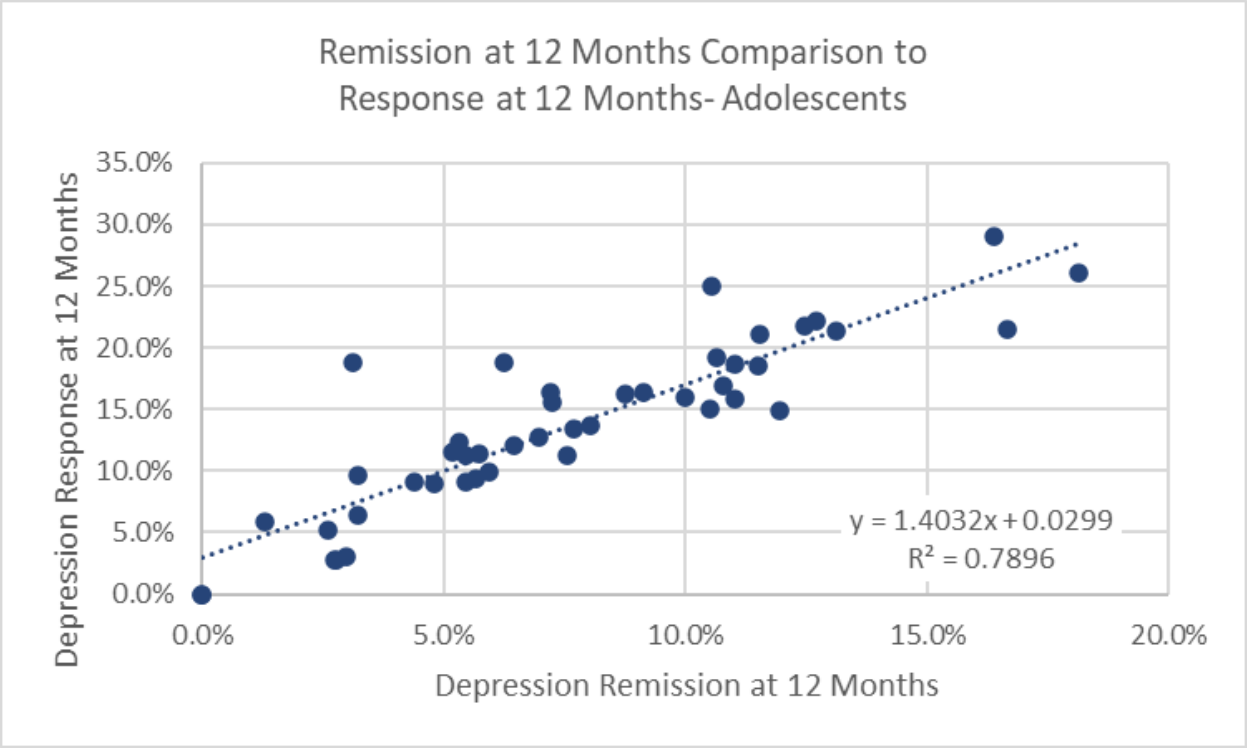


Correlation between Depression Remission at 12 Months and Patients who Achieve Response but Not Remission at 12 Months; 550 clinics and 118, 132 patients R-Squared value of 0.3578

Validity Performance Scores (Correlation)- Adolescents

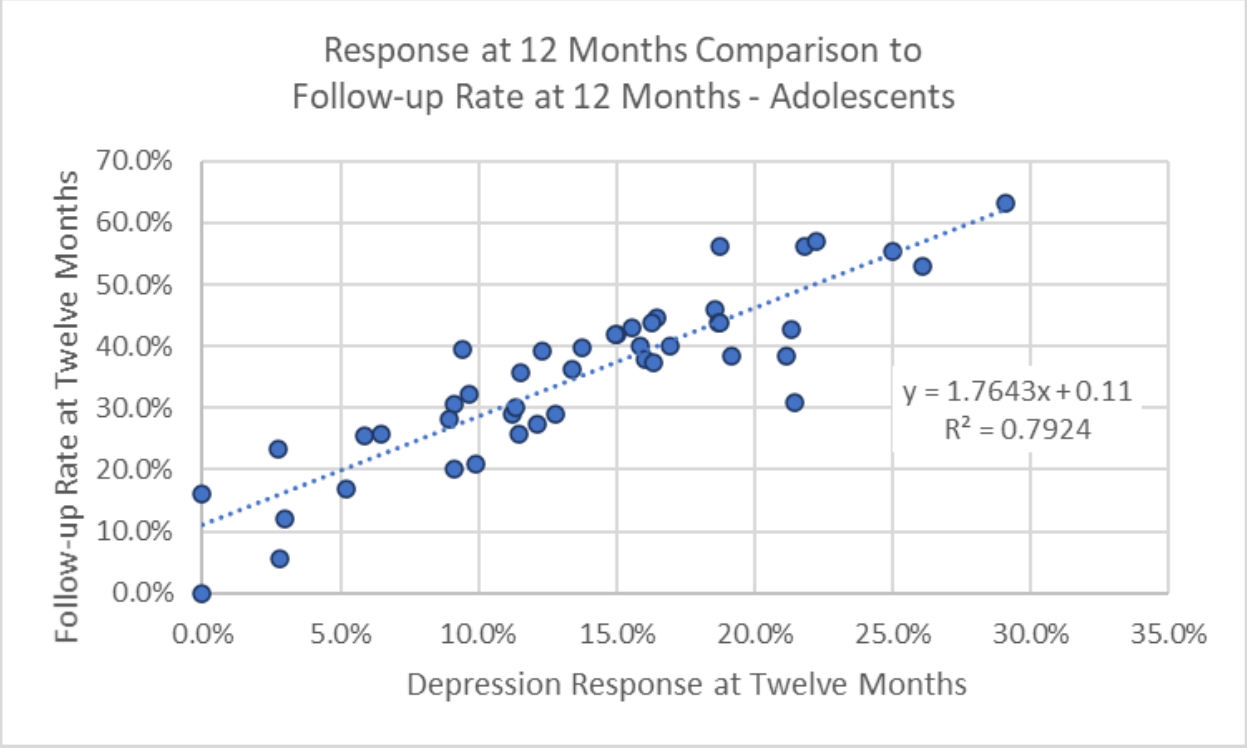
Hypothesis	Description	r-squared
#1	Correlation between depression remission (PHQ-9 < 5) and depression response (PHQ-9 ≥ 50 percent improved from index initial PHQ-9 score)	0.7896
#2	Correlation between depression outcome rates and the rates of follow-up	0.7924
#3	Correlation between patients who achieve remission and those who achieve response but not remission	0.2366

Display of Hypothesis #1 Correlation between Depression Remission at 12 Months and Depression Response at 12 Months



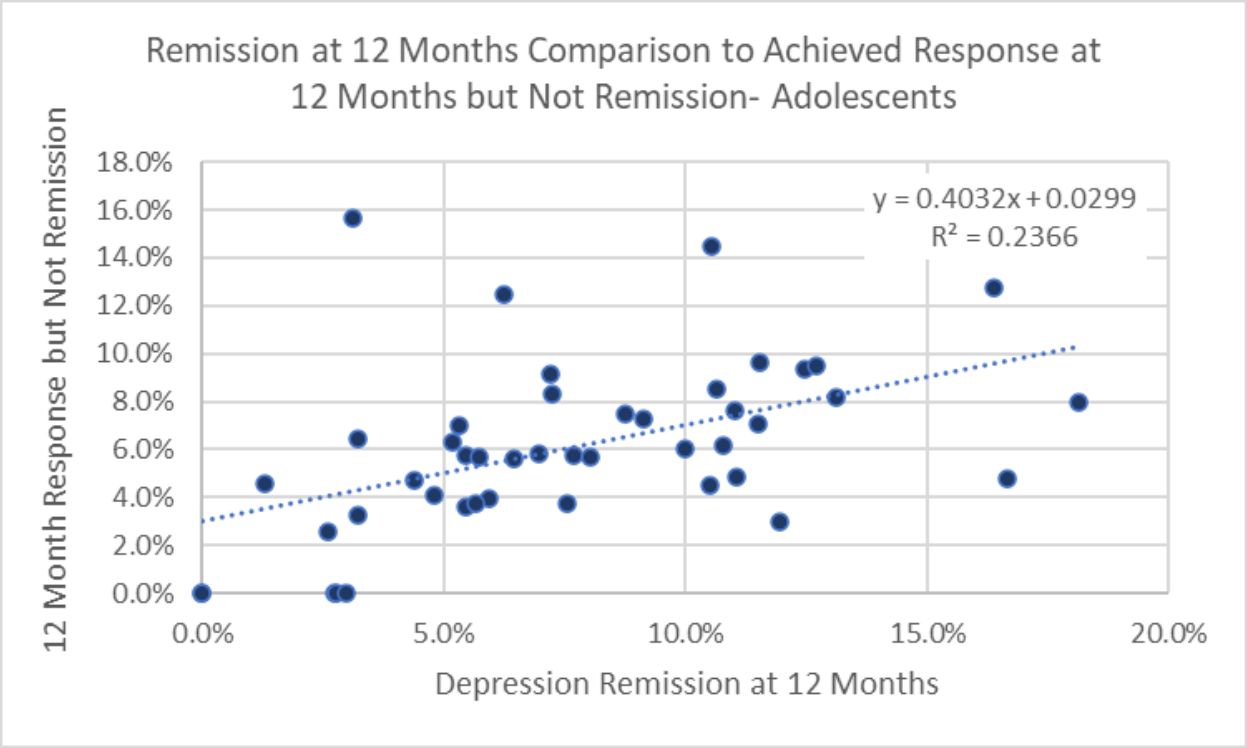
Correlation between Depression Remission at 12 Months and Depression Response at 12 Months; 45 medical groups and 12,115 patients R-Squared value of 0.7896

Display of Hypothesis #2 Correlation between Depression Response at 12 Months and Rate of Depression Follow-up at 12 Months



Correlation between Depression Response at 12 Months and Follow-up Rate at 12 Months; 45 medical groups and 12,115 patients R-Squared value of 0.7924

Display of Hypothesis #3 Correlation between patients who achieve remission and those who achieve response but not remission



Correlation between Depression Remission at 12 Months and Patients who Achieve Response but not Remission ; 45 medical groups and 12,115 patients R-Squared value of 0.2366

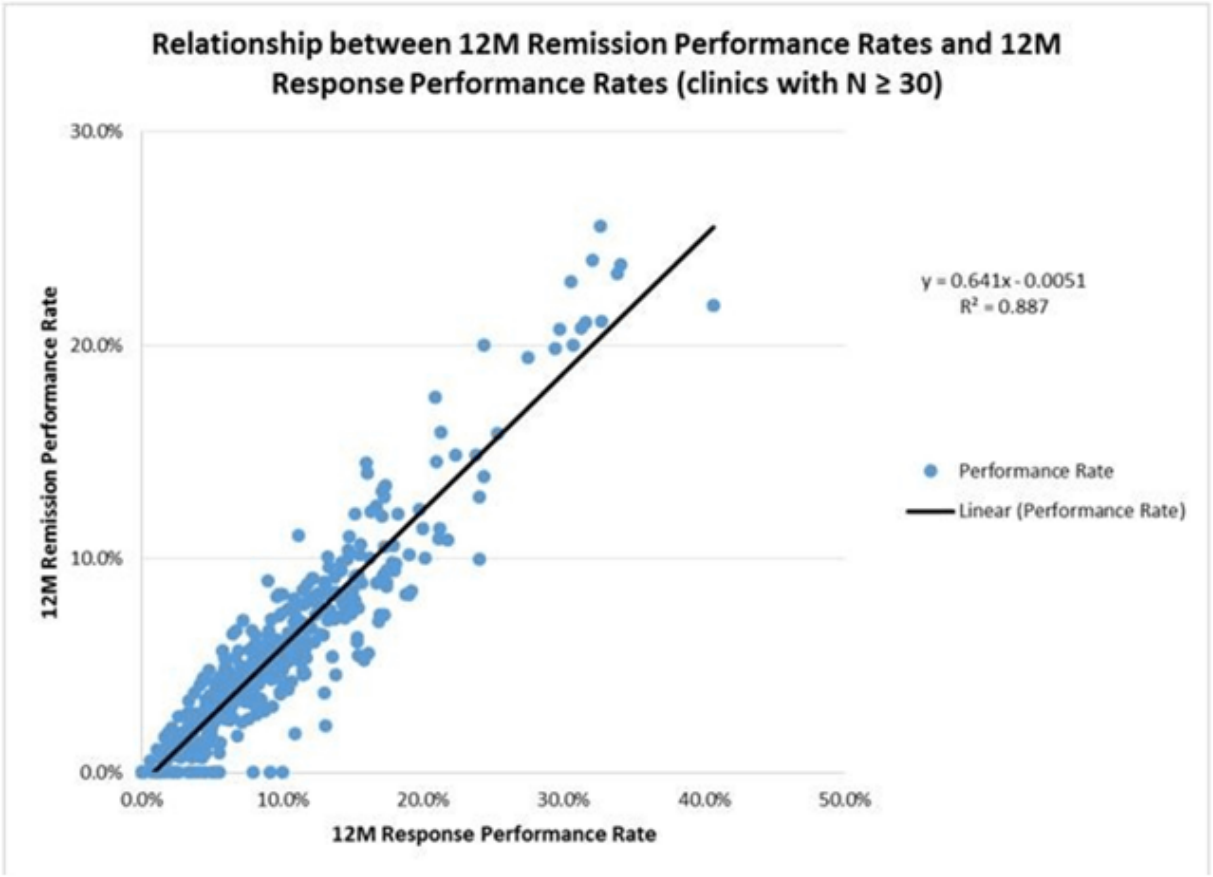
2020 Validation Summary- Data Elements

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

2013 Submission

Validity Performance Score

Correlation between similar depression measures- adults



Validity Performance Score- Correlation between Response and Remission- Adults R-squared = 0.887 Remission = PHQ-9 < 5, Response is ≥ 50% improved

[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 measuring patient outcomes related to depression. There was high compliance with critical data element validity as demonstrated by annual validation audit processes.

The adult stratification demonstrates a high correlation [R squared 0.9051] against a similar measure, confirming the hypothesis that clinics whose patients achieve a response to treatment have more success in achieving depression remission at 12 months. If the coefficient value lies between ± 0.50 and ± 1, then there is said to be a strong correlation.

The adolescent stratification demonstrates a lower correlation value [R squared 0.6026], however this is still in the range of a high correlation. There are potentially two reasons why the correlations between adults and adolescents differ. There are fewer adolescents in the denominator as compared to adults, therefore volume/size may hamper statistical testing, but the second reason may be more explanatory. During the measure redesign process that incorporated adolescents into the measure, the measure development workgroup continually stressed the differences between adults and adolescents (treatments, maturity level, life experiences) and required that the two population's outcomes **always be re**ported separately and never be combined into a single measure.

[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.

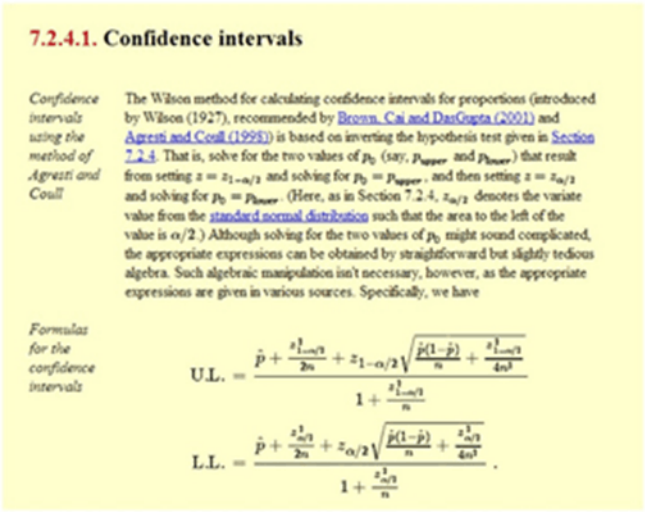
ELIGIBLE POPULATION SPECIFICATIONS The eligible population for each measure is identified by a medical group on behalf of their individual clinics. MNCM's 2019 DDS Data Collection Guides provide technical specifications for the standard definitions of the eligible population, including elements such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

Equation for the Calculation of Confidence Intervals; Wilson Method

This mathematical formula provides the calculation of upper and lower 95%confidence intervals. Equation for the Calculation of Confidence Intervals; Wilson Method

This mathematical formula provides the calculation of upper and lower 95%confidence intervals.



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.

RISK ADJUSTMENT Risk adjustment is a technique used to enable fair comparisons of clinics/medical groups by adjusting for the differences in risk among specific patient groups. MNCM uses an "Actual to Expected" methodology for risk adjustment. This methodology does not alter a clinic/medical group's result; the actual rate remains unchanged. Instead, each clinic/medical group's rate is compared to an "expected rate" for that clinic/medical group based on the specific characteristics of patients seen by the clinic/medical group, compared to the total patient population.

All expected values for DDS measures are calculated using a logistic regression model including the following variables: health insurance product type (commercial, Medicare, Medicaid, uninsured, unknown), patient age, and deprivation index. The deprivation index was added in 2018 and includes ZIP code level average of poverty, public assistance, unemployment, single female with child(ren), and food stamps (SNAP) converted to a single index that is a proxy for overall socioeconomic status.

A population proportions test is used to determine whether there is a statistically significant difference between the expected and actual rates of optimally managed patients attributed to each clinic/medical group. The methodology uses a 95 percent test of significance.

The tables for the risk-adjusted measures include the following information:

- Medical group/clinic name
- Performance
    - "Above Average " = Clinic or medical group's actual rate is significantly above its expected rate
    - "Expected" = Clinic or medical group's actual rate is equivalent to its expected rate
    - "Below Average" = Clinic or medical group's actual rate is significantly below its expected rate
- Patients = Number of patients at a medical group/clinic site that meet the denominator criteria for the measure.
- Actual Rate = Actual percentage of patients meeting criteria (unadjusted rate).
- Expected Rate = Expected percentage of patients meeting criteria based on the clinic's/medical group's mix of patient risk (adjusted rate).
- Actual to Expected Ratio = Actual percentage of patients meeting criteria divided by the expected percentage of patients meeting criteria for the clinic's/medical group's mix of patient risk.

[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.

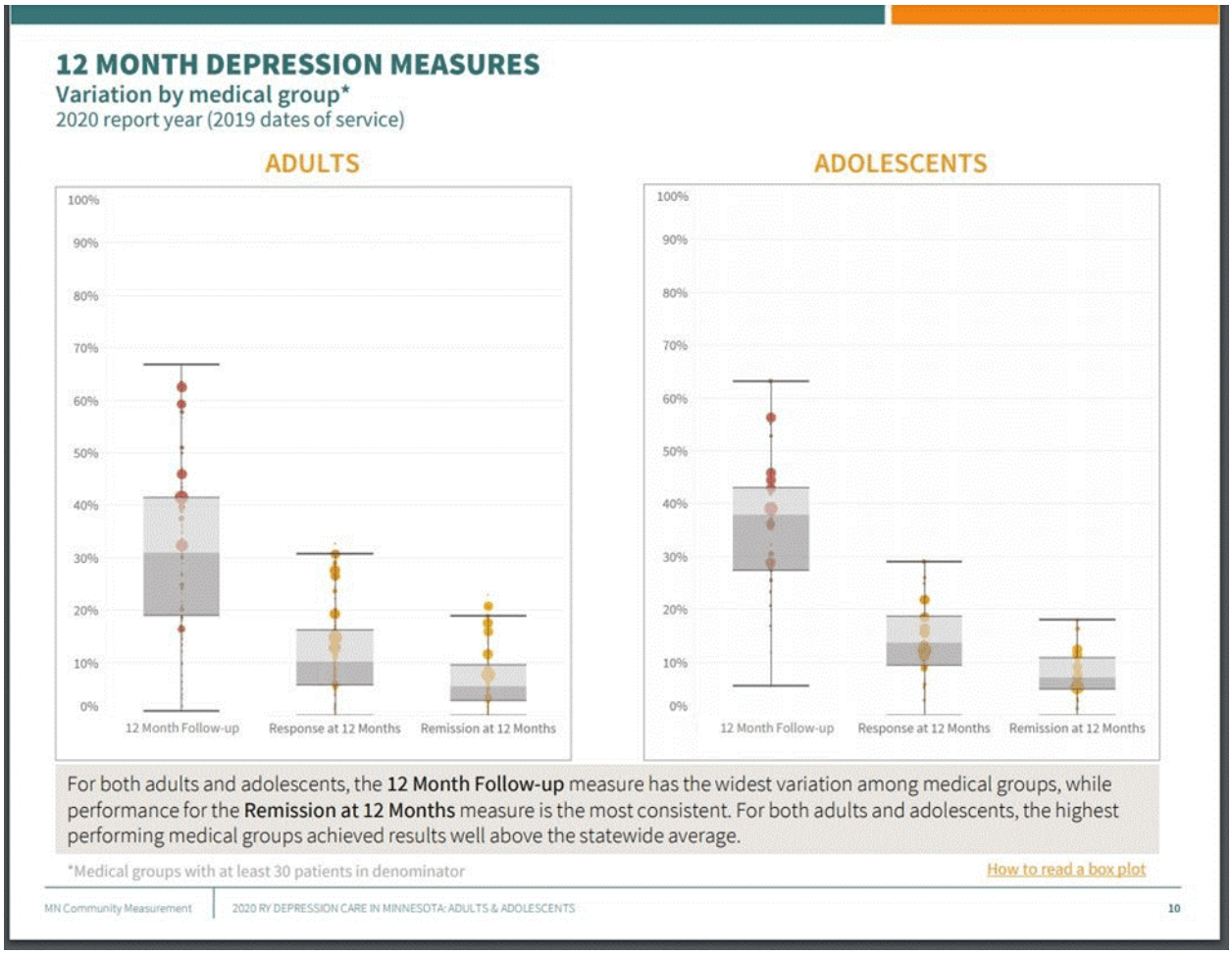
Examples may include number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined.

[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.



MNCM Methods for Identifying Meaningful Differences; Variability Demonstrated by Box Plot Diagram

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

1. Adults 17.0% (range 0% to 32.7%) 120,344 patients from 550 clinics
2. Adolescents 14.5% (range 0% to 29.1%) 11,658 patients from 118 clinics

The box plot diagram demonstrates that many medical groups fall within the upper quartile range. However, the overall rates are low and signal room for improvement.

MEDICAL GROUPS WITH HIGHEST PERFORMANCE

2020 report year (2019 dates of service)
Medical groups with above average performance on at least 50 percent of measures for which they were eligible.

*	Adults	Adults	Adults	Adults	Adults	Adults	Adolescents	Adolescents	Adolescents	Adolescents	Adolescents	Adolescents
MEDICAL GROUP	Six Month Follow-up	Response at Six Months	Remission at Six Months	12 Month Follow-up	Response at 12 Months	Remission at 12 Months	Six Month Follow-up	Response at Six Months	Remission at Six Months	12 Month Follow-up	Response at 12 Months	Remission at 12 Months
Amery Hospital and Clinic	●	●	●	●	●	●	○	○	○	○	○	○
CenterCare Health	○	○	○	○	○	○	●	●	●	●	●	●
Entira Family Clinics	●	●	●	●	●	●	●	●	●	●	●	●
Essential Health	●	●	●	●	●	●	○	○	○	○	○	○
HealthPartners Central Minnesota Clinics	●	●	●	●	●	●	○	○	○	○	○	○
HealthPartners Clinics	●	●	●	●	●	●	○	●	○	●	○	○
Lake Region Healthcare	●	●	●	●	●	●	○	●	○	●	○	●
Lakewood Health System	○	●	○	●	○	●	<	<	<	<	<	<
Mankato Clinic, Ltd.	●	●	●	●	●	●	●	●	●	●	●	●
Olmsted Medical Center	●	●	●	●	●	●	○	○	○	○	○	○
Ortonville Area Health Services	●	●	●	●	●	○	<	<	<	<	<	<
Park Nicollet Health Services	●	●	●	●	●	●	●	●	○	●	○	●
Sanford Health	●	●	●	●	●	●	○	○	○	○	○	○
Westfields Hospital and Clinic	●	●	●	●	●	●	<	<	<	<	<	<

* Cell intentionally left empty

Display of MN Medical Groups Who Achieved Average Performance on at Least 50% of the Eligible Measures

The above image is a display of top medical groups in MN with the highest performance rates, having achieved above average performance on at least 50 percent of the measures. For example, the medical groups Entira Family Clinics and Mankato Clinic achieved above average rates in all measures which is delineated with a gold circle. Measure rates that were average or below the statewide average are designated with an open circle. A carat < indicates that there were too few patients in the denominator (e.g., adolescents) to calculate the measure.

[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]

Measure continues to demonstrate significant opportunity for both maintaining contact with patients with depression (ongoing follow-up) and achieving an outcome of remission. Measure results demonstrate opportunity for improvement in depression outcomes and identify meaningful differences among providers.

[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.

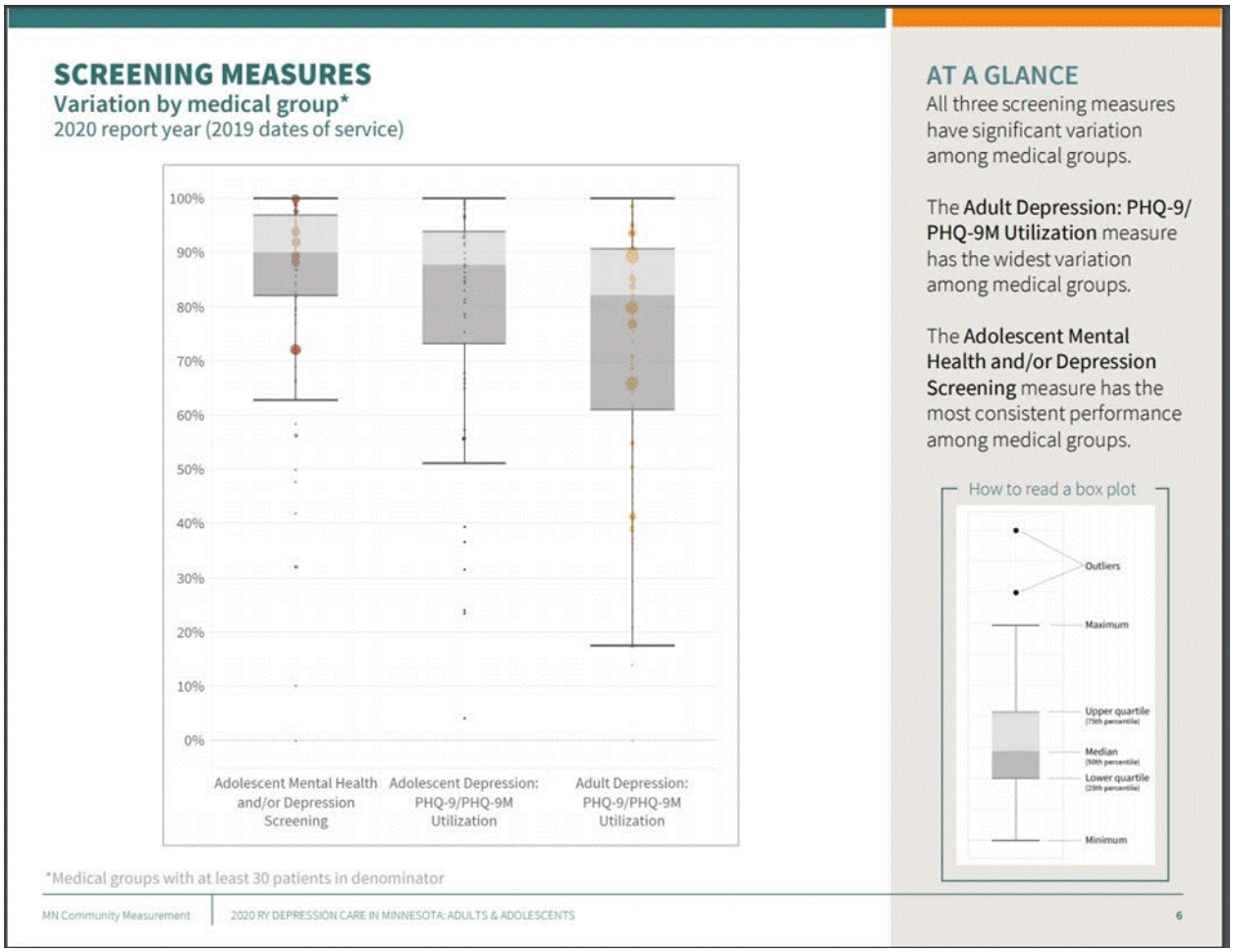
Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]

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 incremental improvements in rates of follow-up PHQ-9 at 12 months, from 17.0% in 2010 to 41.8% in 2019 for adults. Adolescents, a new population for this measure have a 2019 follow-up rate of 38.9%

Missing data, in this case, follow-up PHQ-9 patient reported outcome assessment is not an issue as those patients who are not re-assessed in follow-up remain in the denominator and are treated as if they are not in remission. However, low outcome rates are not solely attributed to lack of follow-up. A portion of patients are still experiencing symptoms of depression and are not in remission. A separate analysis for patients who were assessed with a follow-up PHQ-9 demonstrates that remission was at 24% while significant depression symptoms persisted for 49% of the patients (24% moderate, 15% major and 10% severe)

There is a companion related measure that allows medical groups to understand their use of the PHQ-9/ PHQ-9M tool, NQF # 0712 Depression Utilization of PHQ-9M. This measure reports the rate of tool administration for patients with a diagnosis of depression or dysthymia seen during a four month measurement period.



Companion measure for utilization of the PHQ-9 for patients with major depression/ dysthymia; supports the outcome measures

The image above displays the box plot chart for the companion measure that informs PHQ-9/PHQ-9 usage for patients with a diagnosis of major depression or dysthymia. If there was avoidance of measuring the depression outcome measures of response and remission, a medical group would have a very low rate here as assessing with a PHQ-9/PHQ-

9M tool is required for indexing into the denominator. This diagram shows statewide information and very few outliers with low PHQ-9/PHQ-9M administration rates.

[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 follow-up PHQ-9/ PHQ-9M at twelve months (+/- 60 days) are included in the denominator and treated as if they are not in remission.

[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.09

[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]

When known, exclusions are removed “up-front”, prior to data submission and validated through the denominator certification process as described in 2b.02 and may not be 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.

2021 Submission

With the redesign of this measure to incorporate the adolescent population, the measure development workgroup reviewed all exclusions and enhanced the measure to additionally exclude patients with schizophrenia and pervasive developmental disorder. An updated exclusion analysis was performed in 2020, demonstrating an overall rate of exclusion of 3.45% of 140,099 patients.

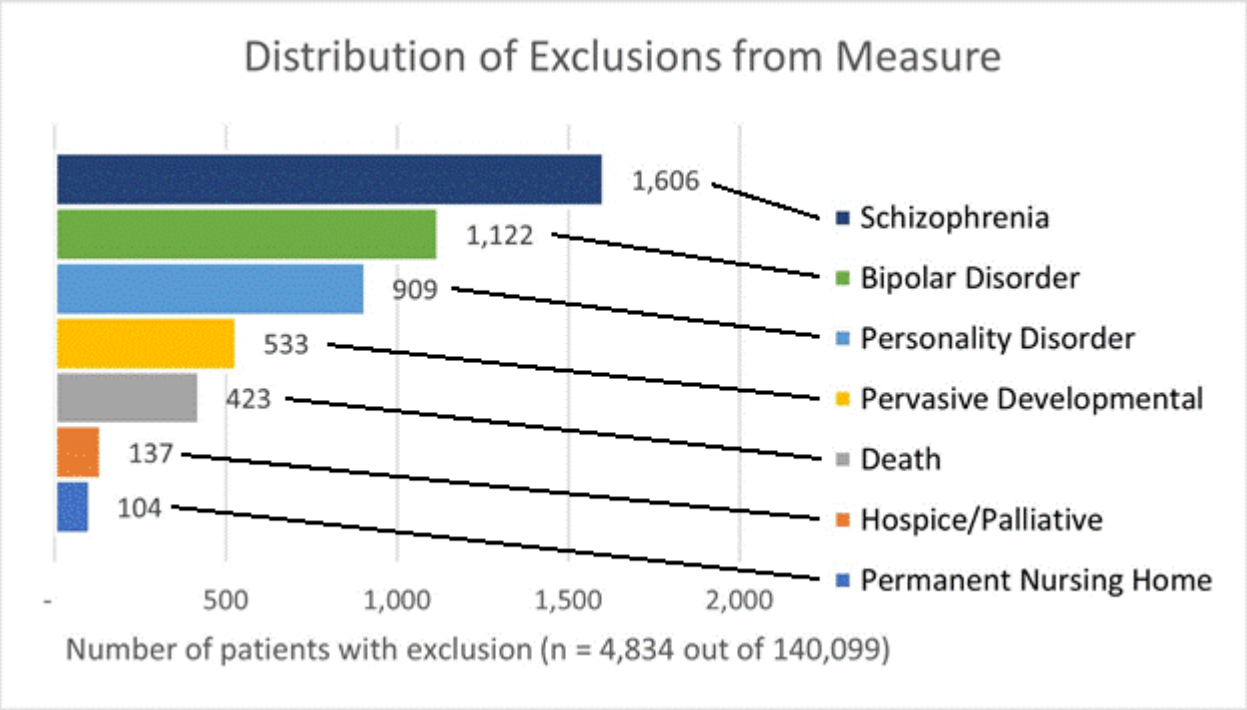
[Response Ends]

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]

2020 Updated Exclusion Analysis



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.

2013 Submission

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.

[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.

2021 Submission

Updated analysis of modifications and additions to exclusions demonstrate continued appropriate clinical indication without reducing the target population. Reliability performance scores for the adult population increased slightly with measure redesign (from 0.900 to 0.9151).

[Response Ends]

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

[Response Begins]

Statistical risk model with risk factors (specify number of risk factors)

[Statistical risk model with risk factors (specify number of risk factors) Please Explain]

Four risk variables: Age, Insurance Type, Depression Severity and Deprivation Index

Patient Age, Patient Insurance Type (Commercial, Medicare, State Public Programs, Uninsured and Unknown Insurance Type), Depression Severity Level at Time of Index (3 levels), and Deprivation Index of Patient Zip Code (percentage of households with SNAP Benefits, Living under Poverty Level, On Public Assistance, Single Female with Children and Percentage of Adults Unemployed.

[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]

MNCM uses Logistic Regression Modeling to create values supporting a method of Indirect Standardization Risk Adjustment, commonly referred to as Expected Value. Indirect standardization does not change the actual performance rates, rather answers the question: "If all providers had this medical group/ clinic's mix of patients, what would the statewide average be?". This method compares the provider's actual performance to the expected rate.

Example Clinic X	**Unadjusted**	**Standardized to Clinic X Patient Mix**
Statewide	39%	32%
Clinic X	35%	35%
Clinic X vs Statewide	Below	Above (Actual : Expected = 1.09

Risk variables used for this measure include age, initial PHQ-9/ PHQ-9M score, insurance product and patient neighborhood deprivation index (based on zip-code). Deprivation index includes use of SNAP benefits, living under the poverty level, unemployed status, public assistance, and single female with children. In MN, the ratio ranges are -6.41 (Red Lake) to +1.42 (Flom) with a mean of zero. "A measure of census-tract neighborhood deprivation is likely a good proxy for a range of individual-level and true area-level constructs relevant to outcomes of interest and feasible to obtain." [National Academies of Sciences, Engineering, and Medicine, 2017: Accounting for Social Risk Factors in Medicare Payment]

2021 Submission

12 Month Response- Adults

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adults**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8846	0.0253	5560.51	<.0001
pt_age	1	0.0112	0.000531	440.8772	<.0001
mdcr	1	-0.2698	0.0251	115.9745	<.0001
mhcp	1	-0.4775	0.0215	495.1298	<.0001
unins	1	-0.4472	0.0473	89.4065	<.0001
undt	1	-0.4456	0.0278	256.3172	<.0001
mod_severe	1	0.0474	0.0175	7.3062	0.0069
severe	1	0.0425	0.0213	3.9953	0.0456
dep_idx	1	0.1376	0.0102	182.9268	<.0001

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Adults

Table of results for data elements selected for the risk stratification model (age, insurance product, severity of depression at index event and deprivation index. All variables have a Chi-squared p value of less than .0001.

12 Month Response- Adolescents

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adolescents**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7016	0.266	6.9551	0.0084
pt_age	1	-0.0671	0.0174	14.8288	0.0001
mdcr	1	-0.5406	0.2661	4.126	0.0422
mhcp	1	-0.2282	0.0642	12.6212	0.0004
unins	1	-0.6443	0.1982	10.5702	0.0011
undt	1	-0.19	0.0879	4.6756	0.0306

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adolescents**	*	*	*	*	*
mod_severe	1	-0.0108	0.0603	0.0323	0.8574
severe	1	0.1504	0.0687	4.8021	0.0284
dep_idx	1	0.1133	0.0405	7.829	0.0051

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Adolescents

Table of results for data elements selected for the risk stratification model (age, insurance product, severity of depression at index event and deprivation index. All variables have a Chi-squared p value of less than .0001.

Definitions for Logistic Model

AIC – This is the Akaike Information Criterion. It is calculated as AIC = -2 Log L + 2((k-1) + s), where k is the number of levels of the dependent variable and s is the number of predictors in the model. AIC is used for the comparison of nonnested models on the same sample. Ultimately, the model with the smallest AIC is considered the best, although the AIC value itself is not meaningful.

SC – This is the Schwarz Criterion. It is defined as – 2 Log L + ((k-1) + s)*log(Σ fi), where fi's are the frequency values of the i^{th} observation, and k and s were defined previously. Like AIC, SC penalizes for the number of predictors in the model and the smallest SC is most desirable and the value itself is not meaningful.

-2 Log L – This is negative two times the log-likelihood. The -2 Log L is used in hypothesis tests for nested models and the value in itself is not meaningful.

Intercept Only – This column refers to the respective criterion statistics with no predictors in the model, i.e., just the response variable.

Intercept and Covariates – This column corresponds to the respective criterion statistics for the fitted model. A fitted model includes all independent variables and the intercept. We can compare the values in this column with the criteria corresponding Intercept Only value to assess model fit/significance.

Test – These are three asymptotically equivalent Chi-Square tests. They test against the null hypothesis that at least one of the predictors' regression coefficient is not equal to zero in the model. The difference between them are where on the log-likelihood function they are evaluated.

Likelihood Ratio – This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model. The LR Chi-Square statistic can be calculated by -2 Log L(null model) – 2 Log L(fitted model) = 231.289-160.236 = 71.05, where L(null model) refers to the Intercept Only model and L(fitted model) refers to the Intercept and Covariates model.

Score – This is the Score Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.

Wald – This is the Wald Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.

Chi-Square, DF and Pr > ChiSq – These are the Chi-Square test statistic, Degrees of Freedom (DF) and associated p-value (PR>ChiSq) corresponding to the specific test that all of the predictors are simultaneously equal to zero. We are testing the probability (PR>ChiSq) of observing a Chi-Square statistic as extreme as, or more so, than the observed one under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model are equal to zero. The DF defines the distribution of the Chi-Square test statistics and is defined by the number of predictors in the model.
Typically, PR>ChiSq is compared to a specified alpha level, our willingness to accept a type I error, which is often set at 0.05 or 0.01. The small p-value from all three tests would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero.

[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]

[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]

Published literature

Internal data analysis

[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]

During the measure development process, the expert panel discusses potential variables for risk adjustment that are important to consider for the measured population, in this case patients with depression. The group decides what clinical variables in addition to the MNCM standard demographic data (gender, age, zip, race/ethnicity, country of origin, primary language, and insurance product) to collect through the data collection and submission process. The potential risk adjustment variables are then evaluated for appropriate inclusion in the model based on a chi square t test value less than 0.05.

Guiding principles for variable selection include the following:

- Conceptual relationship with outcome
- Empirical association with outcome
- Variation across measured entities
- Not confounded with the effect of health care
- Resistant to manipulation or gaming
- Accurate data that can be reliably and feasibly captured
- Contribution of unique variation in the outcome (not redundant)
- Potentially, improvement in risk model
- Potentially, face validity and acceptability

Please refer to the response in question 2b.20 for a description of the Indirect Standardization Risk Adjustment process.

[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]

2021 Submission

12 Month Response- Adults

Updated model includes a variable of deprivation index (dep_idx) (NOTE: Maximum likelihood estimates contained in the output below are the same estimates that appear above in 2b.20)

resp12_a Depression Response at 12 Months- Adults	*	*	*	*
Product Variables	*	*	*	*
The FREQ Procedure	*	*	*	*
prod_nm	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Commercial	55813	46.38	55813	46.38
Medicaid	27384	22.75	83197	69.13
Medicare	19609	16.29	102806	85.43
Self-Paid/Uninsured	4183	3.48	106989	88.9
Undetermined	13355	11.1	120344	100
comm	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	64531	53.62	64531	53.62
1	55813	46.38	120344	100
mdcr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	100735	83.71	100735	83.71
1	19609	16.29	120344	100
mhcp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	92960	77.25	92960	77.25
1	27384	22.75	120344	100
mdcd_unins	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	88777	73.77	88777	73.77
1	31567	26.23	120344	100

resp12_a Depression Response at 12 Months- Adults	*	*	*	*
unins	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	116161	96.52	116161	96.52
1	4183	3.48	120344	100
undt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	106989	88.9	106989	88.9
1	13355	11.1	120344	100

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adults

resp12_a Depression Response at 12 Months- Adults	*	*
Product Variables	*	*
The LOGISTIC Procedure	*	*
Model Information	*	*
Data Set	RA.RESP12_A_PROD_VARS	*
Response Variable	response_12	*
Number of Response Levels	2	*
Model	binary logit	*
Optimization Technique	Fisher's scoring	*
Number of Observations Read	120344	*
Number of Observations Used	120344	*
Response Profile	*	*
Ordered Value	response_12	Total Frequency
11	*	20450
20	*	99894

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adults

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adults**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8846	0.0253	5560.51	<.0001
pt_age	1	0.0112	0.000531	440.8772	<.0001
mdcr	1	-0.2698	0.0251	115.9745	<.0001
mhcp	1	-0.4775	0.0215	495.1298	<.0001
unins	1	-0.4472	0.0473	89.4065	<.0001
undt	1	-0.4456	0.0278	256.3172	<.0001
mod_severe	1	0.0474	0.0175	7.3062	0.0069
severe	1	0.0425	0.0213	3.9953	0.0456
dep_idx	1	0.1376	0.0102	182.9268	<.0001

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adults

Odds Ratio Estimates **Depression Response at 12 Months- Adults**	*	*	*
Effect	Point Estimate	95% Wald Confidence Limits	*
pt_age	1.011	1.01	1.012
mdcr	0.764	0.727	0.802
mhcp	0.62	0.595	0.647
unins	0.639	0.583	0.702
undt	0.64	0.606	0.676
mod_severe	1.049	1.013	1.085
severe	1.043	1.001	1.088
dep_idx	1.148	1.125	1.171

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adults

12 Month Response - Adolescents

resp12_c Depression Response at 12 Months- Adolescents	*	*	*	*
Product Variables	*	*	*	*
The FREQ Procedure	*	*	*	*
prod_nm	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Commercial	6671	57.22	6671	57.22
Medicaid	3173	27.22	9844	84.44
Medicare	159	1.36	10003	85.8
Self-Paid/Uninsured	328	2.81	10331	88.62
Undetermined	1327	11.38	11658	100
comm	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4987	42.78	4987	42.78
1	6671	57.22	11658	100
mdcr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11499	98.64	11499	98.64
1	159	1.36	11658	100
mhcp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8485	72.78	8485	72.78
1	3173	27.22	11658	100
mdcd_unins	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	8157	69.97	8157	69.97
1	3501	30.03	11658	100
unins	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11330	97.19	11330	97.19
1	328	2.81	11658	100
undt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	10331	88.62	10331	88.62
1	1327	11.38	11658	100

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adolescents

resp12_c **Depression Response at 12 Months- Adolescents**	*	*
Product Variables	*	*
The LOGISTIC Procedure	*	*
Model Information	*	*
Data Set	RA.RESP12_C_PROD_VARS	*
Response Variable	response_12	*
Number of Response Levels	2	*
Model	binary logit	*
Optimization Technique	Fisher's scoring	*
Number of Observations Read	11658	*
Number of Observations Used	11658	*
Response Profile	*	*
Ordered Value	response_12	Total Frequency
11	*	1694
20	*	9964

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adolescents

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adolescents**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7016	0.266	6.9551	0.0084
pt_age	1	-0.0671	0.0174	14.8288	0.0001
mdcr	1	-0.5406	0.2661	4.126	0.0422
mhcp	1	-0.2282	0.0642	12.6212	0.0004
unins	1	-0.6443	0.1982	10.5702	0.0011

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adolescents**	*	*	*	*	*
undt	1	-0.19	0.0879	4.6756	0.0306
mod_severe	1	-0.0108	0.0603	0.0323	0.8574
severe	1	0.1504	0.0687	4.8021	0.0284
dep_idx	1	0.1133	0.0405	7.829	0.0051

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adolescents

Odds Ratio Estimates **Depression Response at 12 Months- Adolescents**	*	*	*
Effect	Point Estimate	95% Wald Confidence Limits	*
pt_age	0.935	0.904	0.968
mdcr	0.582	0.346	0.981
mhcp	0.796	0.702	0.903
unins	0.525	0.356	0.774
undt	0.827	0.696	0.982
mod_severe	0.989	0.879	1.113
severe	1.162	1.016	1.33
dep_idx	1.12	1.035	1.212

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Frequency and Logistic Procedures Adolescents

2013 Submission

Original Model Development:

Originally, the depression remission at six months measure (#0710) was tested to determine the appropriate selection of variables using the following method:

The effect of risk adjustment on clinic ranking is examined in three ways. First, the clinic's unadjusted and adjusted quality measures are compared using correlation analysis. Two types of correlation are used, Pearson and Kendall. Pearson's correlation examines the correlation when the measures are treated as continuous measures. A high correlation (close to 1) means that the two measures strongly co-vary, when one is high the other is high. Kendall's correlation examines the similarity between the unadjusted and adjusted quality measure in terms of the similarity in the way clinics are ranked by the measures. Because of the focus of Kendall's correlation on comparing ranks and the interest in the use of clinic quality scores for clinic comparison, Kendall's correlation is likely to be the most useful correlation measure.

The second comparison ranks the clinics into performance rank deciles based on the unadjusted and adjusted scores and then examines how decile rankings based on unadjusted measures compare to decile rankings based on adjusted measures. The third comparison ranks clinics into Poor, Below Average, Average, Above Average, and Excellent categories

using statistical methods that take into account the quality measure's confidence interval which is calculated based on the number of patients each clinic reports. These two methods are compared directly in our accompanying report on the quality deviations ranking approach.

Risk adjustment is necessary only when there is heterogeneity across clinics. There was significant heterogeneity across clinics in insurance product mix ($\chi 2 = 10120$, $p < .001$), patient age ($\chi 2 = 5325$, $p < .001$), gender ($\chi 2 = 1267$, $p < .001$), initial severity ($\chi 2 = 1759$, $p < .001$), and distance to the clinic ($\chi 2 = 35015$, $p < .001$).

Table 1 Effect of Potential Risk Adjusters on Depression

1A Model without SES and Race from Zip Code Data

Category	**Variable**	**Contrast**	**Estimate**	**T-value**	**Odds Ratio**	**Lower 95% CI**	**Upper 95% CI**
Age	Age 18 - 25	66+	-0.62**	-6.66	0.54**	0.45	0.65
Age	Age 26 - 50	66+	-0.68**	-8.70	0.51**	0.44	0.59
Age	Age 51 - 65	66+	-0.66**	-8.48	0.52**	0.45	0.60
Gender	Female	Male	-0.08	-1.88	0.92	0.85	1.00
Depression Severity	Moderate	Severe	0.57**	10.56	1.77**	1.59	1.97
Depression Severity	Moderately Severe	Severe	0.39**	6.84	1.48.**	1.33	1.66
Distance from Clinic	< 5 miles	Same Zip	-0.05	-0.86	0.95	0.85	1.07
Distance from Clinic	5 - 10 miles	Same Zip	-0.09	-1.44	0.92	0.82	1.03
Distance from Clinic	10 - 20 miles	Same Zip	-0.06	-1.03	0.94	0.83	1.06
Distance from Clinic	20+ miles	Same Zip	-0.10	-1.33	0.90	0.78	1.05
Insurance	Medicare	Commercial	-0.48**	-9.72	0.55**	0.48	0.63
Insurance	Medicaid/ MSHO/ Special Needs/ Self-pay/Uninsured	Commercial	-0.59**	-8.83	0.62**	0.56	0.68
Constant	*	*	-1.85	-1.76	*	*	*

* Cell intentionally left empty

** indicates statistical significance

1B Model with SES and Race from Zip Code Data

Category	*	*	*	*
Age	Age 18 - 25	66+	-0.62**	-6.65
Age	Age 26 - 50	66+	-0.68**	-8.67
Age	Age 51 - 65	66+	-0.66**	-8.47
Gender	Female	Male	-0.08	-1.87
Depression Severity	Moderate	Severe	0.57**	10.51
Depression Severity	Moderately Severe	Severe	0.39**	6.80
Distance from Clinic	< 5 miles	Same Zip	-0.05	-0.78
Distance from Clinic	5 - 10 miles	Same Zip	-0.09	-1.45
Distance from Clinic	10 - 20 miles	Same Zip	-0.07	-1.19

* Cell intentionally left empty

[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]

2021 Submission

MNCM staff met with a team of researchers at the University of Minnesota that work with health disparities research to better understand if Race, Ethnicity, Language and Country of Origin (RELO) variables met the criteria of having a conceptual relationship (i.e., race should affect the measure) and were not confounded by the clinic's contribution. The data demonstrates that RELO variables do have an impact to some degree but proving both a conceptual relationship and not being a confounding factor was not a consensus that the MNCM Risk Adjustment Committee could reach. They concluded that geography is what should be considered. Neighborhoods are what truly matter ; an actual neighborhood defined by census block tracks. Neighborhoods appear to incorporate numerous factors that do impact risk adjustment. They include some parts of RELO, but also median income, traditional family wealth, incarceration rates, food, single family homes, safety, truancy, ambient noise level and factors we know to be social determinants of health.

Race, Ethnicity, Language and Country of Origin (RELO) data were not used because of potential implicit bias. For these reasons, the deprivation index was selected as a proxy for social determinants of health.

[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]

2013- The analyses were conducted at the patient level, with patients nested within clinics. Patient characteristics, such as age, gender, initial severity, insurance product, distance between the patient's zip code and the clinic's zip code, and an indicator (fixed effect) for each clinic were included in the model. The patient characteristics measure the relationship between those characteristics and patient outcomes. The clinic indicators measure clinic differences in performance controlling for patient characteristics. The clinic indicators also control for unobserved differences across clinics that may be correlated with the risk adjusters and result in biased estimates of the risk adjustment effects.

The analysis of whether follow-up is correlated with remission was done using Stata's implementation of Heckman's method for correcting for sample selection. The sample selection occurs because remission is observed only for those who are followed up at six months. The Heckman procedure estimates two models: one for follow-up and one for remission. The procedure tests whether follow-up is correlated with remission. The measures included in the follow-up equation are age, gender, initial severity, insurance product, and distance to the clinic. The measures included in the remission equation are age, gender, initial severity, insurance product, and a clinic fixed effect.

A logistic model specification that accounts for the binary nature of remission (depression is a binary outcome - yes/no) is used. Severity at initial diagnosis, age, gender, and insurance product were included as risk adjusters.

2021 Submission Update

The analyses were conducted at the patient level and then rolled up to the clinic level to complete testing at the level of the clinic. Patient characteristics, such as age, initial severity, insurance product, deprivation index and an indicator (fixed effect) for each clinic was included in the model. The patient characteristics measure the relationship between those characteristics and patient outcomes. The clinic indicators measure clinic differences in performance controlling for patient characteristics. The clinic indicators also control for unobserved differences across clinics that may be correlated with the risk adjusters and result in biased estimates of the risk adjustment effects.

The measures included in the follow-up equation are age, initial severity, insurance product, and zip code level deprivation index. The measures included in the remission equation are age, initial severity, insurance product, deprivation index and a clinic fixed effect. A logistic model specification that accounts for the binary nature of remission (depression is a binary outcome – yes/no) is used.

[Response Ends]

2b.27. Provide risk model discrimination statistics.

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

[Response Begins]

2021 Submission

12 Month Response- Adults

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adults**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8846	0.0253	5560.51	<.0001
pt_age	1	0.0112	0.000531	440.8772	<.0001
mdcr	1	-0.2698	0.0251	115.9745	<.0001
mhcp	1	-0.4775	0.0215	495.1298	<.0001
unins	1	-0.4472	0.0473	89.4065	<.0001
undt	1	-0.4456	0.0278	256.3172	<.0001
mod_severe	1	0.0474	0.0175	7.3062	0.0069
severe	1	0.0425	0.0213	3.9953	0.0456
dep_idx	1	0.1376	0.0102	182.9268	<.0001

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Adults

Table of results for data elements selected for the risk stratification model (age, insurance product, severity of depression at index event and deprivation index. All variables have a Chi-squared p value of less than .0001.

12 Month Response- Adolescents

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adolescents**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7016	0.266	6.9551	0.0084
pt_age	1	-0.0671	0.0174	14.8288	0.0001
mdcr	1	-0.5406	0.2661	4.126	0.0422
mhcp	1	-0.2282	0.0642	12.6212	0.0004
unins	1	-0.6443	0.1982	10.5702	0.0011
undt	1	-0.19	0.0879	4.6756	0.0306
mod_severe	1	-0.0108	0.0603	0.0323	0.8574
severe	1	0.1504	0.0687	4.8021	0.0284
dep_idx	1	0.1133	0.0405	7.829	0.0051

*Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Adolescents

Table of results for data elements selected for the risk stratification model (age, insurance product, severity of depression at index event and deprivation index. All variables have a Chi-squared p value of less than .0001.

2013 Submission

2013- Risk adjustment is necessary only when there is heterogeneity across clinics. There was significant heterogeneity across clinics in insurance product mix ($\chi2 = 10120$, $p < .001$), patient age ($\chi2 = 5325$, $p < .001$), gender ($\chi2 = 1267$, $p < .001$), initial severity ($\chi2 = 1759$, $p < .001$), and distance to the clinic ($\chi2 = 35015$, $p < .001$).

[Response Ends]

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

[Response Begins]

2021 Submission

12 Month Response- Adults

Impact of Risk Adjustment on Comparison to Mean	*	*	*	*	*
Clinic Distribution	*	*	*	*	*

Impact of Risk Adjustment on Comparison to Mean	*	*	*	*	*
*	*	Risk Adjusted Comparison	*	*	*
*	*	Below Expected	Expected	Above Expected	*
Unadjusted	Below Average	100^	**78+**	0	**178**
*	Average	**6***	235^	0	241
*	Above Average	0	**46***	85^	131
*	*	106	359	85	550

* Cell intentionally left empty

Two Dimensional Table Displaying the Impact of Risk Adjustment for Individual Clinics

The above table is a two-dimensional display of the impact of risk adjustment. If the 178 clinics that are statistically below the mean before adjustment, 78 of those clinics are considered "Expected" (green cell/+ symbol) once the social and medical factors are considered. Conversely as indicated by blue cells/ * symbol, 6 clinics who were considered average prior to risk adjustment decreased to below expected and 46 clinics who were rated above average were only meeting the expected risk adjusted rate. The gray cells/ ^ symbol represents the number of clinics (majority) whose rating did not change as a result of risk adjustment.

12 Month Response- Adolescents

*	**Adolescent 12 Month Response**	*	*	*	*
*	Clinic Distribution	*	*	*	*
*	*	Risk Adjusted Comparison	*	*	*
*	*	Below Expected	Expected	Above Expected	*
Unadjusted	Below Average	2^	3	0	5
*	Average	0	104^	0	104
*	Above Average	0	**4***	5^	**9**
*	*	2	111	5	118

* Cell intentionally left empty

Two Dimensional Table Displaying the Impact of Risk Adjustment for Individual Clinics

The above table is a two-dimensional display of the impact of risk adjustment. 4 clinics that were rated above average had their ranking changed to expected as a result of risk adjustment (blue cells/ * symbol). The gray cells/ ^ symbol represents the number of clinics (majority) whose rating did not change as a result of risk adjustment.

The design of this risk adjustment is that clinics with higher risk patients are given a lower threshold to mean and the clinics with lower risk patients are given a higher threshold when compared to all other clinics.

2013 Submission

Original Depression Remission at Six Months

Comparison of Unadjusted and Adjusted Decile Ranks (N/Percent of Row)

*	Risk Adjusted Decile Rank	*	*	*	*	*	*	*	*	*	*
Unadjusted Decile Rank	0 to 10%	10% to 20%	20% to 30%	30% to 40%	40% to 50%	50% to 60%	60% to 70%	70% to 80%	80% to 90%	90% to 100%	Total
0 to 10%	21 80.77^	4 15.38+	0 0.00	0 0.00	0 0.00	1 3.85+	0 0.00	0 0.00	0 0.00	0 0.00	26
10% to 20%	2 8.33#	16 66.7^	6 25.00+	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	24
20% to 30%	2 7.14#	0 0.00#	12 42.86^	7 25.00+	6 21.43+	0 0.00	1 3.45+	0 0.00	0 0.00	0 0.00	28
30% to 40%	1 3.45#	5 17.24#	5 17.24#	9 31.03^	6 20.69+	2 6.90+	0 0.00	1 3.45+	0 0.00	0 0.00	29
40% to 50%	0 0.00	1 4.35#	3 11.54#	7 30.43#	7 30.43^	5 21.74+	0 0.00	0 0.00	0 0.00	0 0.00	23
50% to 60%	0 0.00	0 0.00	0 0.00	3 11.54#	6 23.08#	10 38.46^	5 19.23+	2 7.69+	0 0.00	0 0.00	26
60% to 70%	0 0.00	0 0.00	0 0.00	0 0.00	1 3.85#	6 23.08#	10 38.46^	7 26.92+	1 3.85#	1 3.85#	26
70% to 80%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	2 7.69#	8 30.77#	12 46.15^	4 15.38#	0 0.00	26
80% to 90%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	2 7.69#	4 15.38#	17 65.38^	3 11.54+	26
90% to 100%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	4 15.38#	22 84.62^	26
Total	26	26	26	26	26	26	26	26	26	26	260

* cell intentionally left blank
^ no change in rank
+ increase in rank after risk adjustment
# decrease in rank after risk adjustment

Original Model Development; Adults
Depression Remission at 12 Months; Clinic Distribution Before and After Risk Adjustment
2013 Dates of Service
Risk Variables: Product, Severity and Age Band

*	**After Risk Adjustment**	*	*	*
Before Risk Adjustment	Below	Expected	Above	Total
Significantly Below	39^	50+	0+	89
Average	0#	249^	1+	250
Significantly Above	0#	9#	48^	97
*	39	308	49	396
Better	51	13%	*	*
Same	336	85%	*	*
Worse	9	2%	*	*

* cell intentionally left blank
^ no change in rank
+ increase in rank after risk adjustment
# decrease in rank after risk adjustment

Original Statistical Output Decile Rank Change

[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]

2021 Submission

12 Month Response- Adults

Impact of Risk Adjustment on Comparison to Mean	*	*	*	*	*
Clinic Distribution	*	*	*	*	*
*	*	Risk Adjusted Comparison	*	*	*
*	*	Below Expected	Expected	Above Expected	*
Unadjusted	Below Average	100^	**78+**	0	**178**
*	Average	**6***	235^	0	241
*	Above Average	0	**46***	85^	131
*	*	106	359	85	550

* Cell intentionally left empty

Two Dimensional Table Displaying the Impact of Risk Adjustment for Individual Clinics

The above table is a two-dimensional display of the impact of risk adjustment. If the 178 clinics that are statistically below the mean before adjustment, 78 of those clinics are considered “Expected” (green cell/+ symbol) once the social and medical factors are considered. Conversely as indicated by blue cells/ * symbol, 6 clinics who were considered average prior to risk adjustment decreased to below expected and 46 clinics who were rated above average were only meeting the expected risk adjusted rate. The gray cells/ ^ symbol represents the number of clinics (majority) whose rating did not change as a result of risk adjustment.

12 Month Response- Adolescents

*	**Adolescent 12 Month Response**	*	*	*	*
*	Clinic Distribution	*	*	*	*
*	*	Risk Adjusted Comparison	*	*	*
*	*	Below Expected	Expected	Above Expected	*
Unadjusted	Below Average	2^	3	0	5
*	Average	0	104^	0	104
*	Above Average	0	**4***	5^	**9**
*	*	2	111	5	118

* Cell intentionally left empty

Two Dimensional Table Displaying the Impact of Risk Adjustment for Individual Clinics

The above table is a two-dimensional display of the impact of risk adjustment. 4 clinics that were rated above average had their ranking changed to expected as a result of risk adjustment (blue cells/ * symbol). The gray cells/ ^ symbol represents the number of clinics (majority) whose rating did not change as a result of risk adjustment.

2013 Submission

Original Depression Remission at Six Months

*	Risk Adjusted Decile Rank	*	*	*	*	*	*	*	*	*	*
Unadjusted Decile Rank	0 to 10%	10% to 20%	20% to 30%	30% to 40%	40% to 50%	50% to 60%	60% to 70%	70% to 80%	80% to 90%	90% to 100%	Total
0 to 10%	21 80.77^	4 15.38+	0 0.00	0 0.00	0 0.00	1 3.85+	0 0.00	0 0.00	0 0.00	0 0.00	26
10% to 20%	2 8.33#	16 66.7^	6 25.00+	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	24
20% to 30%	2 7.14#	0 0.00#	12 42.86^	7 25.00+	6 21.43+	0 0.00	1 3.45+	0 0.00	0 0.00	0 0.00	28
30% to 40%	1 3.45#	5 17.24#	5 17.24#	9 31.03^	6 20.69+	2 6.90+	0 0.00	1 3.45+	0 0.00	0 0.00	29
40% to 50%	0 0.00	1 4.35#	3 11.54#	7 30.43#	7 30.43^	5 21.74+	0 0.00	0 0.00	0 0.00	0 0.00	23
50% to 60%	0 0.00	0 0.00	0 0.00	3 11.54#	6 23.08#	10 38.46^	5 19.23+	2 7.69+	0 0.00	0 0.00	26
60% to 70%	0 0.00	0 0.00	0 0.00	0 0.00	1 3.85#	6 23.08#	10 38.46^	7 26.92+	1 3.85#	1 3.85#	26
70% to 80%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	2 7.69#	8 30.77#	12 46.15^	4 15.38#	0 0.00	26
80% to 90%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	2 7.69#	4 15.38#	17 65.38^	3 11.54+	26
90% to 100%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	4 15.38#	22 84.62^	26
Total	26	26	26	26	26	26	26	26	26	26	260

* cell intentionally left blank
^ no change in rank
+ increase in rank after risk adjustment
# decrease in rank after risk adjustment
Original Model Development; Adults

Depression Remission at 12 Months; Clinic Distribution Before and After Risk Adjustment
2013 Dates of Service
Risk Variables: Product, Severity and Age Band

*	**After Risk Adjustment**	*	*	*
Before Risk Adjustment	Below	Expected	Above	Total
Significantly Below	39^	50+	0+	89
Average	0#	249^	1+	250
Significantly Above	0#	9#	48^	97
*	39	308	49	396
Better	51	13%	*	*
Same	336	85%	*	*
Worse	9	2%	*	*

* cell intentionally left blank
^ no change in rank
+ increase in rank after risk adjustment
# decrease in rank after risk adjustment

[Response Ends]

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

[Response Begins]

2021 Submission

12 Month Response- Adults

Impact of Risk Adjustment on Comparison to Mean	*	*	*	*	*
Clinic Distribution	*	*	*	*	*
*	*	Risk Adjusted Comparison	*	*	*
*	*	Below Expected	Expected	Above Expected	*
Unadjusted	Below Average	100^	**78+**	0	**178**
*	Average	**6***	235^	0	241
*	Above Average	0	**46***	85^	131
*	*	106	359	85	550

* Cell intentionally left empty

Two Dimensional Table Displaying the Impact of Risk Adjustment for Individual Clinics

The above table is a two-dimensional display of the impact of risk adjustment. If the 178 clinics that are statistically below the mean before adjustment, 78 of those clinics are considered “Expected” (green cell/+ symbol) once the social and medical factors are considered. Conversely as indicated by blue cells/ * symbol, 6 clinics who were considered average prior to risk adjustment decreased to below expected and 46 clinics who were rated above average were only meeting the expected risk adjusted rate. The gray cells/ ^ symbol represents the number of clinics (majority) whose rating did not change as a result of risk adjustment.

12 Month Response- Adolescents

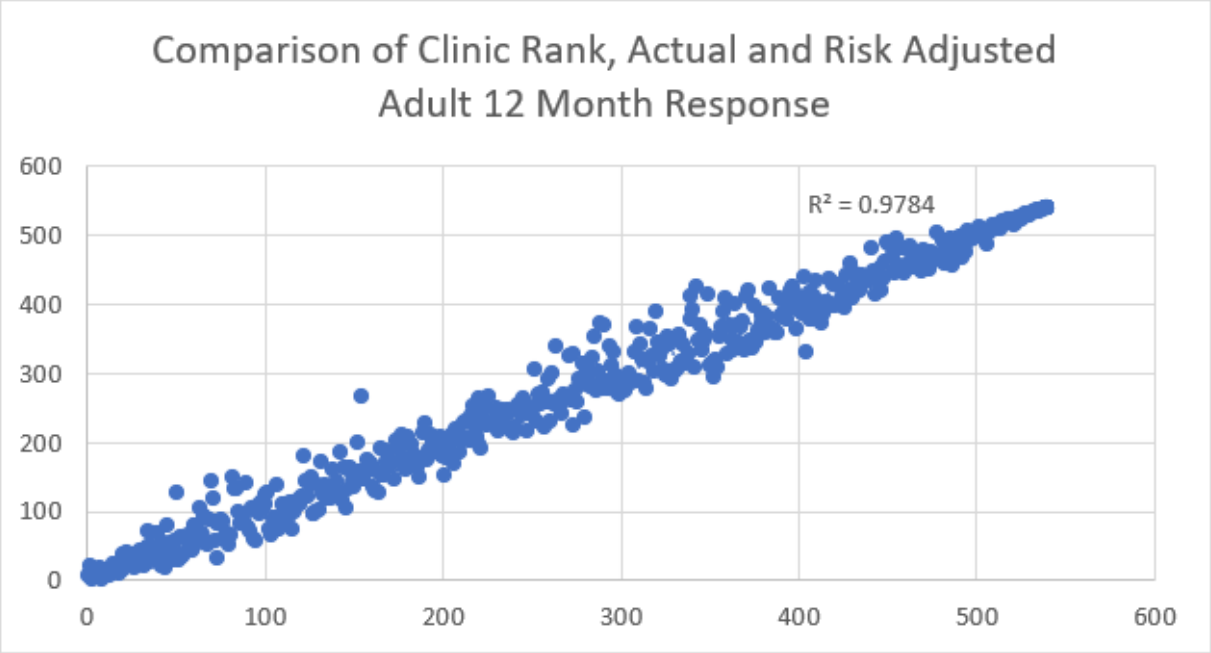
*	**Adolescent 12 Month Response**	*	*	*	*
*	Clinic Distribution	*	*	*	*
*	*	Risk Adjusted Comparison	*	*	*
*	*	Below Expected	Expected	Above Expected	*
Unadjusted	Below Average	2^	3	0	5
*	Average	0	104^	0	104
*	Above Average	0	**4***	5^	**9**

*	**Adolescent 12 Month Response**	*	*	*	*
*	*	2	111	5	118

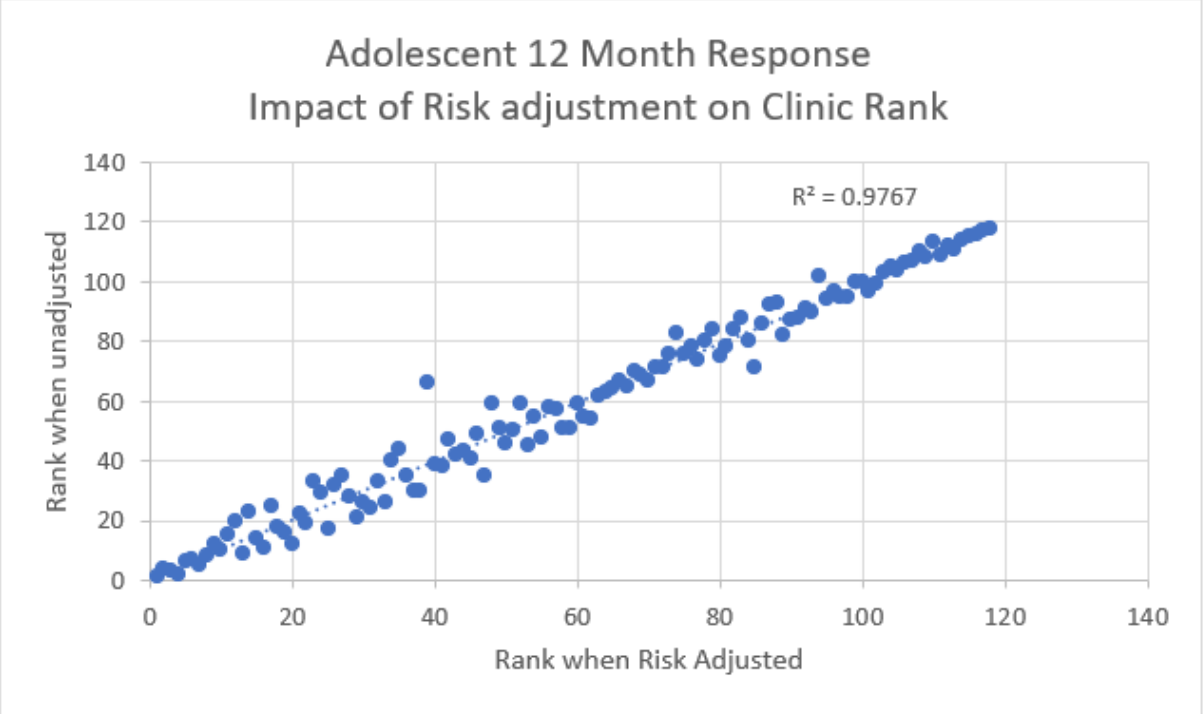
* Cell intentionally left empty

Two Dimensional Table Displaying the Impact of Risk Adjustment for Individual Clinics

The above table is a two-dimensional display of the impact of risk adjustment. 4 clinics that were rated above average had their ranking changed to expected as a result of risk adjustment (blue cells/ * symbol). The gray cells/ ^ symbol represents the number of clinics (majority) whose rating did not change as a result of risk adjustment.



Comparison of Clinic Rank; Actual to Expected Risk adjusted Rate



Comparison of Clinic Rank; Actual to Expected Risk adjusted Rate

Above is a comparison of the clinic ranking when unadjusted (vertical) and with Risk Adjustment (horizontal), The R Squared of the trend line is .9784 for adults and 0.9767 for the adolescent population, proving a high correlation between unadjusted and adjusted values. Risk adjustment should not greatly alter the results but instead fine tune at the edges for the clinics with unusual patient risk.

*	**Adult 12 Month response**	*	*	*	*
*	Average Risk Score at Clinic Level	*	*	*	*
No Risk Adjustment	*	Below Expected	Expected	Above Expected	*
*	Below Average	1.03	**1.11+**	0.00	**1.11**
*	Average	**0.93***	1.00	0.00	1.01
*	Above Average	0.00	**0.97***	0.98	0.97
*	*	1.05	1.03	0.98	1.00

* Cell intentionally left empty

Average Risk Score Areas of Change

The average risk level for clinics that are originally listed as below average is 1.11 (green cell/ + symbol)(state average is 1.0), the below average clinics that are reevaluated as "expected" have a higher risk level 1.11, than the clinics that remained below average (1.03). Clinics who ranking changed from average to below expected or above average to expected are shaded in blue/ *symbol. This is how the model is supposed to work in that there are not radical shifts based on risk variables which could indicate measure or risk model instability.

Does this risk adjustment model make sense for the clinics who have an expected value that is higher or lower?

Evaluated the ten highest risk clinics and the ten lowest risk clinics, does it make sense based on the type of clinic, clinic characteristics and socioeconomic/ demographic in which the clinic is located? Clinic characteristics demonstrated results as expected; clinics with lower socioeconomic values had more patients at risk.

Clinics with highest risk patient population	**Clinic Characteristic Supporting Risk Variables**
Advanced Medical Clinic, Inc.	Clinic focus is for economically limited patients, high portion of uninsured
Indian Health Board of Minneapolis	Focus is for inner city Native American population, which usually is economically limited
Native American Community Clinic	Focus is for inner city Native American population, which usually is economically limited
NorthPoint Health & Wellness Center	Federally Qualified Healthcare Center (FHCQ), serving culturally diverse and economically limited population
Planned Parenthood Minnesota, North Dakota, South Dakota - Duluth	Young patient population, which is more transient
Ramsey County Mental Health Center	Clinic focus is for economically limited patients, high portion of uninsured
Stark Clinic- Northside	Rule 29 mental health clinic, located in a culturally diverse and economically limited geographic location
Stark Clinic-York	Rule 29 mental health clinic, located in a culturally diverse and economically limited geographic location

Clinics with highest risk patient population	Clinic Characteristic Supporting Risk Variables
University of Minnesota Physicians - Broadway Family Medicine Clinic	Clinic focus is for economically limited patients, high portion of uninsured
West Side Community Health Services - McDonough Homes Clinic	Clinic focus is for economically limited patients, high portion of uninsured
Clinics with lowest risk patient population	**Clinic Characteristic Supporting Risk Variables**
Allina Health - Abbott Northwestern General Medicine Associates - Edina	located in a high income suburb
Allina Health - Sharpe Dillon Cockson & Associates	located in a high income suburb
Glencoe Regional Health Services - Lester Prairie	rural location with 98% white population, high school grad or greater 89%
HealthPartners - Mahtomedi Clinic	located in a high income suburb
M Health Fairview Clinic Edina	located in a high income suburb
Northwest Family Physicians - Rogers	rural location with 91% white population, high school grad or greater 96%
Park Nicollet Clinic - Shorewood	located in a high income suburb
Park Nicollet Clinic - St. Louis Park Internal Medicine	located in a high income suburb
Richfield Medical Group	located in a moderate to high income suburb
Sanford Sioux Falls Internal Medicine Clinic	Mid-sized city with 85% white population, high school grad or greater 92%

2013 Submission

We tested the overall correlation between the unadjusted and risk adjusted depression measure using two methods, a Pearson correlation and a Kendall's Tau correlation. In both cases, the value 1 represents a perfect correlation and the value 0 represents a complete lack of correlation between unadjusted and adjusted measures. The Pearson correlation compares the risk adjusted and unadjusted clinic depression values and is .95 which shows a very strong correlation between the unadjusted and adjusted depression measure. The Kendall's Tau correlation compares unadjusted and adjusted rank order of clinics and was .81. This is still a strong correlation, but not as strong as the .95 correlation between risk-adjusted and unadjusted clinic values.

[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]

2021 Submission

12 Month Response-Adults

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adults**	-	-	-	-	-
Compared to Patients with Commercial Insurance and Moderate Depression	-	-	-	-	-

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adults**	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8846	0.0253	5560.51	<.0001
pt_age	1	0.0112	0.000531	440.8772	<.0001
mdcr	1	-0.2698	0.0251	115.9745	<.0001
mhcp	1	-0.4775	0.0215	495.1298	<.0001
unins	1	-0.4472	0.0473	89.4065	<.0001
undt	1	-0.4456	0.0278	256.3172	<.0001
mod_severe	1	0.0474	0.0175	7.3062	0.0069
severe	1	0.0425	0.0213	3.9953	0.0456
dep_idx	1	0.1376	0.0102	182.9268	<.0001

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Adults

Table of results for data elements selected for the risk stratification model (age, insurance product, severity of depression at index event and deprivation index. All variables have a Chi-squared p value of less than .0001.

12 Month Response Adolescents

Analysis of Maximum Likelihood Estimates **Depression Response at 12 Months- Adolescents**	*	*	*	*	*
Compared to Patients with Commercial Insurance and Moderate Depression	*	*	*	*	*
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7016	0.266	6.9551	0.0084
pt_age	1	-0.0671	0.0174	14.8288	0.0001
mdcr	1	-0.5406	0.2661	4.126	0.0422
mhcp	1	-0.2282	0.0642	12.6212	0.0004
unins	1	-0.6443	0.1982	10.5702	0.0011
undt	1	-0.19	0.0879	4.6756	0.0306
mod_severe	1	-0.0108	0.0603	0.0323	0.8574
severe	1	0.1504	0.0687	4.8021	0.0284
dep_idx	1	0.1133	0.0405	7.829	0.0051

* Cell intentionally left empty

SAS Statistical Software Output Analysis of Variables Selected for Risk Adjustment; Adolescents

Table of results for data elements selected for the risk stratification model (age, insurance product, severity of depression at index event and deprivation index. All variables have a Chi-squared p value of less than .0001.

Our analysis of risk adjustment factors for the measure indicates that age, depression severity at diagnosis, insurance provider variables (MSHO, Medicaid, and Medicare) and zip code-based deprivation index are related to depression remission.

Tests of significance at .01%

- All tested factors remain significant.

After analyzing the entire Depression suite of measures, it was reconfirmed that Age, Product, Severity Levels and Deprivation Index are important and significant factors in the outcome, are present at the initial patient encounter, are beyond the control of the provider and all variables are already being collected so no additional provider burden is required.

2013 Submission

2013- Tests of significance at .01%

- MHCP and Uninsured are significant factors; Medicare and Commercial are not significant from each other
- All four age ranges are significant from each other
- Severity is a significant factor

Results at 6 Months follow the same pattern as with the twelve month measure. It appears that the main reason for the change in remission from 8.1 to 5.7 is the change in follow up. For the patients who did return for the visit; they had the same level of remission at 12 months (25.2%) compared to 6 months (24.9%)

[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]

No additional statistical testing.

[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]

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]

N/A

[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]

This measure is captured in electronic health records, but is not currently specified as an e-CQM. Groups can successfully extract the stored PRO tool information from their EHR independent of the Measure Authoring Tool (MAT). Because all data elements can successfully be extracted from EHR systems and with the implementation of MNCM's warehouse-based data collection methodology, this measure would be considered a digital measure.

Several years ago, in discussions with CMS staff, it is our understanding that the MAT could not support the programmatic math needed to calculate 50% or greater reduction from the initial PHQ-9 score.

[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. Use

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]

Several mechanisms for publicly reporting this measure are in place. Consumer-facing public website MN HealthScores is located at https://www.mnhealthscores.org/ rates (including actual, expected and health score rating) are available for every clinic in MN and surrounding border communities. Measure is published as part of the MNCM Annual Health Care Quality Report, Annual Disparities by Insurance Type and Disparities by Race, Ethnicity, Language, Country of Origin and the focus of several issue briefs. https://mncm.org/reports/#community-reports

Payment Program

[Payment Program Please Explain]

This measure was selected for inclusion as a quality metric to measure health outcomes for CMS' CMMI Innovation Model Kidney Care First. MNCM worked with CMS staff in understanding the measure and determined that application of improving depression outcomes would be appropriate within patient populations of chronic kidney disease and end stage renal failure. MNCM shared statewide results to help determine potential benchmarks for this program.

https://www.cms.gov/newsroom/fact-sheets/kidney-care-first-kcf-and-comprehensive-kidney-care-contracting-ckcc-models

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]

Public reporting

[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]

[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 all medical groups who submit data for this state-wide measure via several options:

- Preliminary measure rates are provided immediately after file upload to HIPAA secure, password protected data portal
- A two-week review process is conducted to allow groups to review and potentially appeal prior to public reporting of rates
- Rates are reported by medical group and clinic level on public website MN Healthscores at www.mnhealthscores.org/
- Additionally, rates including all historical rates can be obtained from the MNCM data portal (pass-word protected)

[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]

Currently, data is collected once per year and results are provided on an annual basis. See question 4a.05 for the process and list of multiple mechanisms for receiving results and providing feedback.

MNCM provides recorded webinars for each measure or measure set that provides education for measure specification (denominator, numerator, exclusions) measure calculation and understanding results.

Education and explanation are also included in our hard copy reports. The annual Health Care Quality Report provides descriptive information along with the results for each measure plus appendices for guidelines for comparing measures over time, data sources and data collection, and methodology (attribution, weighting, rate calculation, risk adjustment). http://mncm.org/reports-and-websites/reports-and-data/health-care-quality-report/

[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]

A similar measure is included in CMS' MIPS and e-CQM program; feedback and comments are provided through the JIRA system. Responses to questions, concerns and suggestions are required to be completed within 48 hours of the question submission. Several clarifications of the measure specifications have occurred as a result of this process. Because this measure is part of a suite of outcome measures for the same denominator of patients (harmonized), any measure changes would be applied across all measures.

[Response Ends]

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

[Response Begins]

MNCM periodically conducts a survey of medical groups in which all clinics in the state are invited to participate and provide feedback. There are structured questions asking the users about measure value and burden.

2018 Medical Group Survey

To what degree does your medical group find value in the measure? (n = 124)

High Value 17.7% (22)

Moderate Value 37.9% (47)

- 56% rated the measure as moderate or high value

  How easy or difficult is it to obtain the data needed for DDS submission for this measure? (n = 124)

  Very Easy 11.3% (14)

  Easy 37.1% (46)

  Difficult 29.8% (37)

  Very Difficult 21..8% (27)

MNCM anticipates a significant drop in burden when the PIPE data collection system is implemented for all groups in MN by year end 2023.

[Response Ends]

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

[Response Begins]

The MNCM Measurement and Reporting Committee, a multi-stakeholder committee of the MNCM Board of Directors, reviews and recommends approval of the slate of measures for public reporting on an annual basis.

https://mncmsecure.org/website/MARC/Slate%20of%20MNCM%20Measures%20for%202021%20Reporting_FINAL.pdf

[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]

As indicated in question spma.o2: Briefly describe any important changes to the measure specifications since the last measure update and provide a rationale.

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; widen the follow-up assessment window; 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. Widen the follow-up assessment window to +/- 60 days for all populations and all response and remission measures

* Six-month measure's assessment window expands from 5 to 7 months to 4 to 8 months

* Twelve-month measure's 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. 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.

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: 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.

5. 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.

Please refer to the data dictionary (sp.11) for the summary of redesign activities and changes to value sets or https://helpdesk.mncm.org/helpdesk/KB/View/22742768--depression-changes-and-rationale

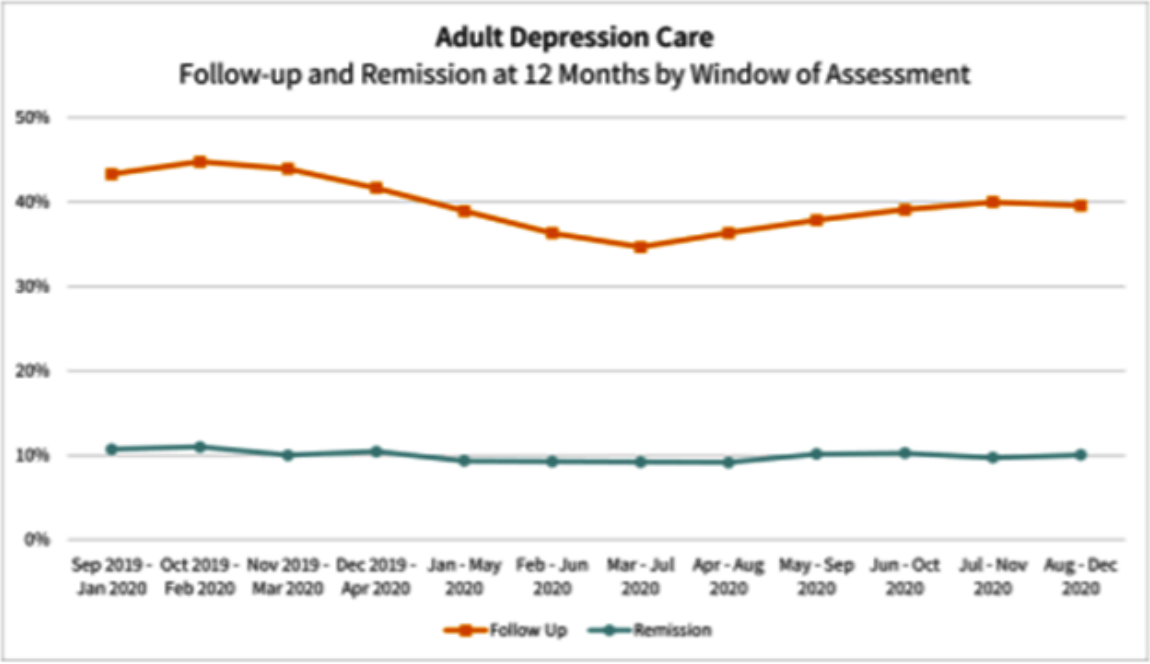
[Response Ends]

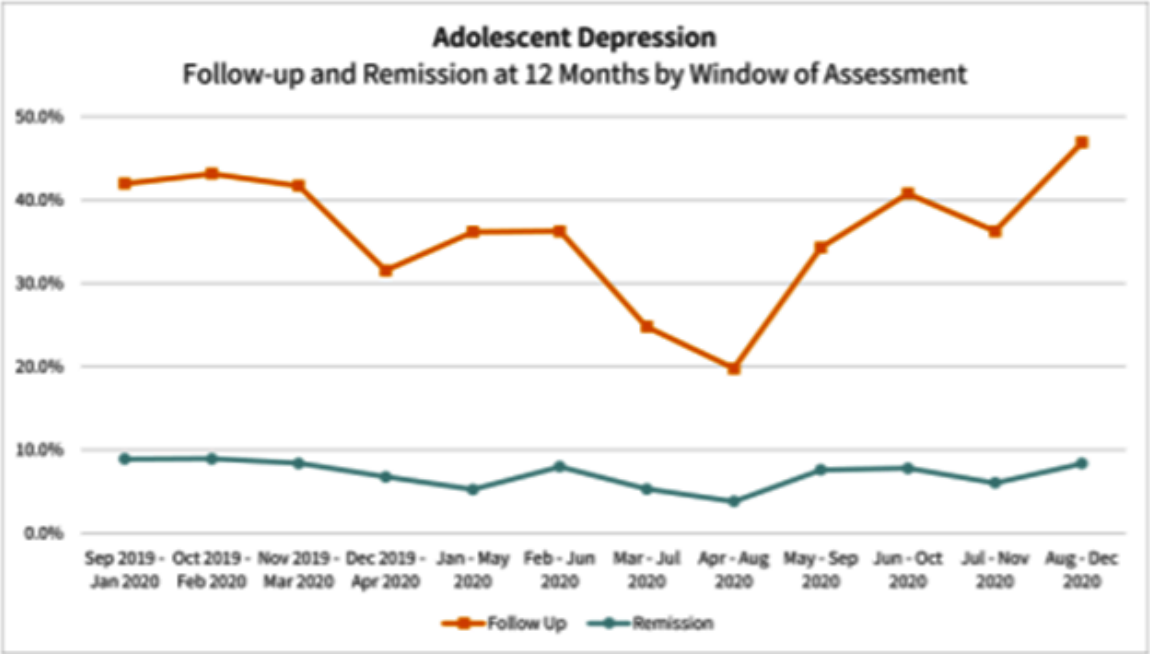
4b. Usability

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]

It is not possible to provide trending information for this measure over time due to recent measure redesign that expanded the assessment window. One measure that we can track for a two year period of time is medical groups' ability to follow-up with their patients. However, this was confounded by the extensive changes in the care delivery system as a result of the pandemic. This related measure, Depression Remission at 12 months demonstrates the comparative line (blue) of the rate of follow-up at 12 months; both of these measures are calculated on the same denominator of patients and require the same follow-up within the measure construct:





Display of Month-to-Month Trend of Remission and Follow-up Rates during COVID-19 Pandemic

[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]

0712: Depression Assessment with PHQ-9/ PHQ-9M

1884: Depression Response at Six Months- Progress Towards Remission

0711: Depression Remission at Six Months

0710e: Depression Remission at Twelve Months

[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]

N/A

[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]

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]

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]

N/A

[Response Ends]