**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** 0712

**Measure Title**: Depression Utilization of PHQ-9

**Date of Submission**: 12/2/2016

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process |
| Efficiency | Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** 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**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** 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 Note: Text in black is from the previous measure testing form. The original testing for the adult population stratification is included and indicated in black text. New testing data for the adolescent population stratification has been added and this new content is in redtext.

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

Please Note: eMeasure testing in cooperationwith CMS subcontractors (Telligen, Mathematica) against CMS test deck completed in April 2014 for continued inclusion in Meaningful Use (MU2)

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

This measure is in full implementation with submission of data from all primary care and behavioral health (psychiatry) clinics in Minnesota. MNCM receives patient counts via a HIPAA secure data portal, so each year data is available for reliability and validity testing on a large population. For this measure, due to its relationship/ paired with longitudinal outcome measures of remission and response, no sampling is allowed and the full population of eligible patients, regardless of payer, is included.

MNCM’s measure development workgroup and reporting body (Measurement and Reporting Committee recently (October 2016) approved updates to the measure that incorporate adolescents. This prompted an ad-hoc review by the NQF Behavioral Health Standing Committee. The adult population stratification is complete with full implementation and data submission from all primary care and behavioral health (psychiatry) clinics in Minnesota. Measure testing for the adult population was submitted in 2014, was deemed adequate and does not require updating. Updates for validity performance score and risk adjustment model statistics are provided.

For the adolescent population stratification, MNCM is working collaboratively with the National Committee for Quality Assurance (NCQA), which explored the use of the measure in adolescents with funding through its AHRQ-CMS Pediatric Quality Measurement Program Center of Excellence. MNCM’s implementation plan for patients with index dates of 1/1/2018 prohibits us from supplying testing information for the adolescent stratification at this time, however NQCA does have testing data for adolescents. We present here the NCQA-conducted testing results to support addition of the adolescent population to the measure. NCQA tested this measure in the adolescent population with three testing sites: two integrated delivery systems and one network of community health centers. The measure was tested using data extracts from the EHR at each test site. We used an initial set of specifications to work with each site to explore patterns of care and methods for identifying patients and extracting the necessary data. Then we obtained from each site an electronic extract, which was used to test alternative versions of the measure specifications. We also obtained manually abstracted data of the electronic record for a subset of charts at two sites to support reliability testing.

**1.3. What are the dates of the data used in testing**? Adult patients with dates of service 10/1/2012 to 1/31/2013 reported in 2013, Adolescents patients age 12 to 17 with dates of service 1/1/2013 to 12/31/2013.

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

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

661 clinic sites were included for this measurement period. 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 who care for adult patients with depression.

For the adolescent population, testing sites were two integrated delivery systems and one network of community health centers that met the following participation criteria: had established clinical workflows for using the PHQ-9 or PHQ-9M, used searchable coded fields for documenting PHQ results in electronic medical records, and had at least 500 adolescents who had a diagnosis of depression in 2012. The sites were from different geographic regions in the U.S. and served urban and rural populations.

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

Adults: 146,135 patients 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.

Adolescents: 3,394 adolescent patients were included in the testing of this measure. Patients met the following inclusion criteria: 1) age 12 to 17 years as of January 1, 2013, and 2) at least one face-to-face visit for a diagnosis of depression during the measure year (January 1, 2013 through December 31, 2013). Adolescents with bipolar, psychotic, autism spectrum, and personality disorders were excluded. No sampling was used: all patients that met the inclusion criteria were included in the testing.

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

There are no differences for the adult stratification for testing.

There are no differences for the adolescent stratification for testing.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

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

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

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 ration 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/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/)]

In addition to the adults and elderly, the PHQ-9 has been validated in 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 [Adult stratification]:

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: σ2 = (α β) / (α + β + 1)(α + β)2
* plug this variance value into the reliability equation: σ2 / (σ2 + (p(1 – p)/n))
  + p = rate



* + n = number of eligible patients
* Determine reliability rate for each clinic.
* Average the reliability rate over all clinics.

Reliability = 0.987

Reliability Testing – Depression PHQ-9 Utilization

Reportable medical groups ( ≥ 30 patients)

* α = 1.3292
* β = 0.9405
* σ2 (provider to provider variance) = 0.0742
* average reliability = 0.987

Population counts submitted by medical groups via HIPAA secure data portal for measure calculation are validated through a process of denominator certification, where groups indicate all criteria and query code used for data submission, which is reviewed and approved by MNCM staff prior to data submission. This insures that all groups are applying denominator criteria correctly lending to consistency and comparability of data and measures.

Additionally, through validation audits performed on a sample of medical groups every year, the presence and accuracy of the PHQ-9 values within the patient’s medical records are assessed.

Critical data elements are audited (demographic information, diagnosis of depression, PHQ-9 date, PHQ-9 score) against the medical record utilizing NCQA’s “8 and 30” audit process and we require a 90% passing rate to allow data to be used for measure calculation and public reporting.

In 2014, for the depression measures, MNCM audited 104 medical groups; 73% of those submitting data. 99% of groups achieved the desired 90% data accuracy; 1 group decided to not correct/ resubmit data and the rates were not included in measure calculation/ public reporting.

Please refer to the validation audit results in section 2b2.2.

Reliability Testing for Adolescents:

In our reliability testing for adolescents, we obtained manually abstracted data of the electronic record for a sample of charts at one site to support reliability testing. We used parallel-forms reliability testing to evaluate the agreement between the electronic extract and the manual review in identifying the numerator, denominator and exclusions for the measure. Agreement was measured using the kappa statistic (a measure of agreement adjusted for agreement that can occur by chance).

Results for parallel-forms reliability testing to evaluate the agreement between the electronic extract and the manual review in identifying the numerator, denominator and exclusions for the measure:

Agreement for the denominators and numerators (before and after exclusions) was high (Kappa of at least 0.73). The exclusions showed low agreement using the Kappa statistic; however, the lowest percentage agreement was 79% for time period 1 for the monitoring measure.

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

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.

PROM-PM [Adult stratification]

* Reliability score = 0.987
* High level of individual data element validity with 99% of groups achieving a 90% accuracy (submitted data matched medical record data)

[Adolescent stratification]

| **Time Period** | **Variable** | **Number of Comparisons** | **Kappa** | **% Agreement** |
| --- | --- | --- | --- | --- |
| Time 1  (1/1/13 – 4/31/13) | Denominator Before Exclusions | 136 | 0.82 | 0.93 |
| Exclusions | 34 | 0.38 | 0.79 |
| Numerator Before Exclusions | 34 | 0.93 | 0.97 |
| Numerator After Exclusions | 24 | 1.00 | 1.00 |
| Time 2  (5/1/13 – 8/31/13) | Denominator Before Exclusions | 136 | 0.77 | 0.90 |
| Exclusions | 34 | 0.45 | 0.88 |
| Numerator Before Exclusions | 34 | 0.92 | 0.97 |
| Numerator After Exclusions | 28 | 0.90 | 0.96 |
| Time 3  (9/1/13 – 12/31/13) | Denominator Before Exclusions | 136 | 0.73 | 0.91 |
| Exclusions | 21 | 0.62 | 0.90 |
| Numerator Before Exclusions | 21 | 0.90 | 0.95 |
| Numerator After Exclusions | 17 | 0.88 | 0.94 |

Denominator before exclusions (Calculated among all patients in age range for the measure)

Exclusions (Calculated among patients found by both sources to qualify for the denominator before exclusions)

Numerator before exclusions (Calculated among patients found by both sources to qualify for the denominator before exclusions)

Numerator after exclusions (Calculated among patients found by both sources to qualify for the denominator after exclusions)

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

The 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 PROM-PM measure [Adult]:

In terms of understanding reliability in detecting signal to noise, a reliability score of 0.70 or greater is considered acceptable for drawing conclusions about groups. This data analysis, along with precise specifications and excellent validation results of critical data elements, demonstrates this measure construct to be reliable and detect meaningful differences among provider groups.

Reliability Testing for Adolescents:

The Kappa statistic has the following interpretation:

0.00 Poor

0.01 – 0.20 Slight

0.21 – 0.40 Fair

0.41 – 0.60 Moderate

0.61 – 0.80 Substantial

0.81 – 0.99 Almost perfect

The Kappa of .73 or higher for all elements except the exclusions suggests the measure can be reliably extracted from electronic health records. While the Kappa for exclusions was .38, .45, and .62 for time periods 1, 2, and 3 respectively, (interpreted as fair, moderate, substantial), agreement between the electronic extract and the manual abstraction was 79% or higher for exclusions. The low Kappa statistic is in part due to the low prevalence of exclusions.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

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 ration 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/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/)]

Validity of the PROM-PM:

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.

Denominator certification 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-9 Codes used
* Attestation for position of depression ICD-9 codes
* 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.

Common areas of correction in denominator for this measure included missing query code, incorrect dates listed or incomplete attestation. All were corrected prior to data submission.

This measure is based on population counts submitted by each medical group. Methods for identifying the population and directions for counting patients are outlined in a detailed data collection guide located at <http://mncm.org/wp-content/uploads/2014/01/Depression_Care_Measures_DDS_2014-Final-12.19.2013.pdf>

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.

Through validation audits performed on a sample of medical groups every year, the presence and accuracy of the PHQ-9 values within the patient’s medical records are assessed.

In 2014, for the depression measures, MNCM audited 104 medical groups; 73% of those submitting data. 77% (80/ 104) passed the initial audit, 23% (24/80) required a correction plan and all re-submitted their data and passed the audit with > 90% accuracy. Types of discrepancies noted on audit included PHQ-9 score obtained but not submitted, incorrect PHQ-9 score, diagnosis code or date of birth.

99% of groups achieved the desired 90% data accuracy; 1 group decided to not correct/ resubmit data and the rates were not included in measure calculation/ public reporting.

Method of Assessing Face Validity for Expanding the Measure to the Adolescent Population:

The health‐plan level of this measure was assessed for use in the HEDIS Health Plan Measure Set. As part of this process, NCQA assessed the face validity of the measure using its comprehensive consensus development HEDIS process. NCQA staff shared the measure concepts, supporting evidence and field test results with its standing Behavioral Health Measurement Advisory Panel, Technical Measurement Advisory Panel and additional panels. We posted the measures for Public Comment, a 30‐day period of review that allowed interested parties to offer feedback about the measure. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations.

Step 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work‐up format. The work‐up is vetted by NCQA’s Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel

(TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

Step 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work‐up that includes a testing proposal and (2)

Collaborate with health plans to conduct field‐tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

Step 3: Public Comment is a 30‐day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures. NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public

Comment measures. New measures and changes to existing measures approved by the CPM and NCQA’s Board of Directors will be included in the next HEDIS year and reported as first‐year measures.

Step 4: First‐year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s State of

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

Staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA’s Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

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.
* 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).

PROM-PM [Adult]

* 99% of groups achieved data element accuracy > 90% when submitted data was compared to medical record data (EHR or paper) of the patient

Results of Face Validity Assessment for the Adolescent Population:

Step 1: This measure was adapted for the adolescent population from the existing Minnesota Community Measurement measure. NCQA and numerous expert panels worked together in 2013 and 2014 to identify the most appropriate method for assessing depression outcome among the adolescent patient population. Across the multiple expert panels that reviewed the measure, all panels concluded this measure was specified appropriately to identify depression remission in adolescents.

Step 2: The measure was field‐tested for the adolescent population in 2013 and 2014. It was first posted for a public comment period in October 2014. After reviewing these initial public comment results along with field test results, the CPM recommended to send the health-plan level version of the measure to the HEDIS public comment period with a majority vote in January 2016.

Step 3: The measure was released for HEDIS Public Comment in February 2016 prior to publication in HEDIS. This measure was rated a high priority by many commenters. The CPM recommended moving this measure to first year data collection by a majority vote in May 2016.

Step 4: The measure was introduced in HEDIS 2017. Organizations will voluntarily report this measure in the first year (2017) and the results will be analyzed for public reporting in the following year.

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The 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 PROM-PM measure [Adult]:

High compliance with critical data element validity as demonstrated by annual validation audit processes.

Interpretation of Systematic Assessment of face validity: The expert panels consulted showed good agreement that the measure as specified will accurately differentiate quality across entities. Additionally, this measure was rated as a high priority measure by the expert panels and by those who responded to the public comment. Our interpretation of these results is that this measure has sufficient face validity.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)  
Exclusions for this process measure paired with outcome measures of depression remission and response are harmonized (match exclusions for the outcome measures). Rationale for exclusions are of a clinical nature where expectations for outcomes may be different due to life expectancy (nursing home residents, hospice/ palliative care, death) or co-morbid diagnoses that may emerge after initial impression/ diagnosis of a depressive disorder (bipolar or personality disorder).

Also need a mechanism to exclude bipolar disorder patients who frequently also have diagnosis of major depression despite this being a departure from best coding practices.

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

When known, exclusions are removed “up-front”, prior to data submission and validated through the denominator certification process as described in 2b2.2 and these exclusions are not available for analysis.

Exclusions for the paired outcome measures:

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.

Adolescents:

Of the 7,860 patients who met the initial criteria to be included in the sample across the three testing sites, 896 (11.4%) met the exclusion criteria of having a diagnosis of any of the following: bipolar disorder, psychotic disorder, autism spectrum disorder, or personality disorder.

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Exclusions, especially the allowance for bipolar or personality disorder that can emerge after initial diagnosis of major depression, are key for this measure and appropriate. Overall, exclusions do not limit or reduce the desired target population of patients with major depression or dysthymia.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of categories **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities**.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

**2b4.4. What were the statistical results of the analyses used to select risk factors?**

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics (e.g., c-statistic, R-squared):**

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:**

**2b4.9. Results of Risk Stratification Analysis**:

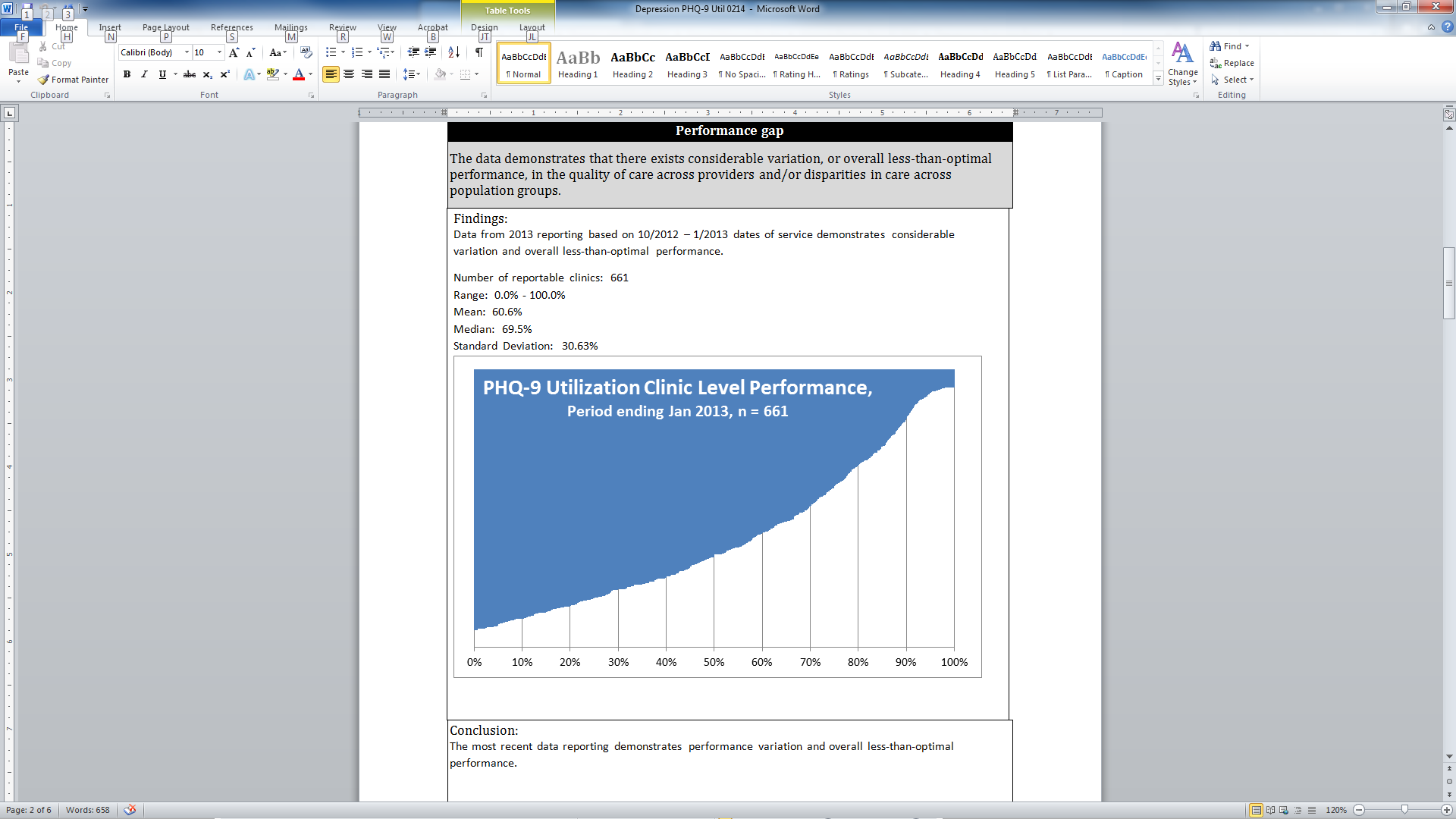
**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

Measure continues to demonstrate significant opportunity for improvement in terms of both administration of the tool to patients with depression on a frequent basis and supports its paired outcome measures by maintaining contact with the patients with depression.



Denominator/ Patients: 146,135  
Numerator: 100,678  
Statewide Average: 65.6%  
  
95% Confidence Interval: 65.3% to 65.8%  
  
Range: 1.1% to 100%  
Median: 63.3%  
Standard Deviation: 26.7%  
  
Distribution Rate Range (# of Clinics)

|  |  |
| --- | --- |
| 0.0 to 9.9% | 68 |
| 10 to 19.9% | 29 |
| 20 to 19.9% | 39 |
| 30 to 39.9% | 28 |
| 40 to 49.9% | 48 |
| 50 to 59.9% | 56 |
| 60 to 69.9% | 58 |
| 70 to 79.9% | 100 |
| 80 to 89.9% | 104 |
| 90 to 100% | 131 |
|  | 661 |

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Please see graphics in 2b5.1

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Measure identifies both opportunity for improvement in depression outcomes and identifies meaningful differences among providers.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

Please note: This outcome measure is paired with a process measure (NQF# 0712 Depression Utilization of the PHQ-9 tool) to allow understanding of the actual administration of the PRO tool within the population of patients who have major depression or dysthymia. NQF# 0712 is rate reflected by the number of patients with major depression or dysthymia who were given at least one PHQ-9 PRO over all patients with major depression or dysthymia who had a visit within a four month measurement period. The process measure adds value/ understanding and promotes frequent use of the PRO. It is not necessary to display the results together.

Though 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 self-advocate and require processes and systems in place for maintaining contact. MN has made small incremental improvements in rates of follow-up PHQ-9 at six and twelve months.

Missing data, in this case PHQ-9 patient reported outcome assessment administered to a patient with depression is not an issue as those patients who are seen and not assessed in the measurement period remain in the denominator.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Patients who are seen and not assessed with a PHQ-9 during the four month measurement period included in the denominator.

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Missing data is not an issue for this measure as constructed; please see discussion in 2b7.1