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

**Measure Number** (*if previously endorsed*)**:** 0711

**Measure Title**: Depression Remission at Six Months

**Date of Submission**: 11/30/2016

**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. The adult population has new validity performance scores (correlations) and this is indicated in red text. New testing data for the adolescent population stratification has been added and this new content is in redtext. Basically, all new information added to the previous testing form (where possible) is indicated in red text.

Spoke with NQF staff and at this time it is acceptable to use the existing testing template with additions in red text and acceptable to exceed the 20 page limit.

**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 |

**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 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.

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 age 18 and older with an index date 7/1/2011 to 6/30/2012 reported in 2013. Adolescents patients age 12 to 17 with an index date 1/1/2012 to 6/30/2013.

**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*)

Participation in the adult measure stratification (age 18 and older) is required by Minnesota health reform legislation- the MN Department of Health’s Statewide Quality Reporting and Measurement System. 403 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*)

86,166 adult 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.

Patients Included in Testing for Adolescent Population:

Across the three testing site, 3,361 adolescent patients were included. Patients met the following inclusion criteria: 1) age 12 to 17 years as of June 30, 2013, 2) at least one face-to-face visit during the study intake period (January 1, 2012 through June 30, 2013), and 3) a diagnosis of depression during the study intake period. 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 = 0.868

Reportable clinics (≥ 30 patients)

α = 1.4768

β = 23.6315

σ2 (provider to provider variance) = 0.00212

average reliability = 0.868

Critical data elements are assessed on a sample of medical groups every year. We audit the critical data elements (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.

In our reliability testing for adolescents, we obtained manually abstracted data of the electronic record for a sample of charts at two sites (n= 46) to support reliability testing. With data from two sites we calculated inter-rater reliability to assess consistency of information obtained on critical data elements used to calculate the measure between manual abstractors. With data from one site we used parallel-forms reliability testing to evaluate the extent to which performance on the quality measure generated automatically by an EHR system was concordant with the performance calculated through manual review of the medical record by trained abstractors. Agreement was measured using the kappa statistic (a measure of agreement adjusted for agreement that can occur by chance).

**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.
* PHQ-9M only slight modification of original tool with developer’s permission

PROM-PM (Adult stratification)

* Reliability score = 0.868
* High level of individual data element validity with 99% of groups achieving a 90% accuracy (submitted data matched medical record data)

Results of Reliability Testing for Adolescents:

Inter-rater agreement among abstractors across the two sites was high with a Kappa of 0.85 or higher for all data elements except for the exclusion data elements.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Number of Comparisons** | **Kappa** | **Lower CI** | **Upper CI** | **% Agreement** |
| Has initial PHQ-9 - Yes/No (within +/- 30 days of visit for depression) | 46 | 1.00 | 1.00 | 1.00 | 1.00 |
| Has/Does not have a PHQ-9 score >9 within +/- 30 days of visit for depression | 46 | 1.00 | 1.00 | 1.00 | 1.00 |
| Follow-Up PHQ-9 - Yes/No (within 4-8 months of first elevated PHQ-9 score) | 46 | 0.95 | 0.86 | 1.00 | 0.98 |
| Any Exclusion Diagnosis | 46 | 0.38 | -0.15 | 0.91 | 0.93 |

Comparison of the electronic extract with manual review for the data elements demonstrated high agreement across sites (Kappa of 1.00) for the “Had follow-up PHQ-9” data element. Percent agreement was 98% across the two sites for the “Had initial PHQ-9” and “Had score >9” despite the Kappa of 0.0. This occurred in a few places where one source showed all patients with one value and then the other source showed one patient with a second value so that one discordant pair results in a Kappa of 0.0, despite 98% agreement. A high level of agreement (92%) was also seen for the exclusion diagnosis data element.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Number of Comparisons** | **Kappa** | **Lower CI** | **Upper CI** | **% Agreement** |
| Has initial PHQ-9 - Yes/No (within +/- 30 days of visit for depression) | 64 | 0.00 | 0.00 | 0.00 | 0.98 |
| Has/Does not have a PHQ-9 score >9 within +/- 30 days of visit for depression | 64 | 0.00 | 0.00 | 0.00 | 0.98 |
| Follow-Up PHQ-9 - Yes/No (within 4-8 months of first elevated PHQ-9 score) | 64 | 1.00 | 1.00 | 1.00 | 1.00 |
| Any Exclusion Diagnosis | 190 | 0.48 | 0.26 | 0.70 | 0.92 |

**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:

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 .85 or higher for all data elements except the exclusions suggests the measure can be reliably extracted from electronic health records. While the overall Kappa for "any exclusion" was 0.38 (interpreted as fair), agreement between abstractors was 92% or higher for all exclusion data elements and "any exclusion". 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:

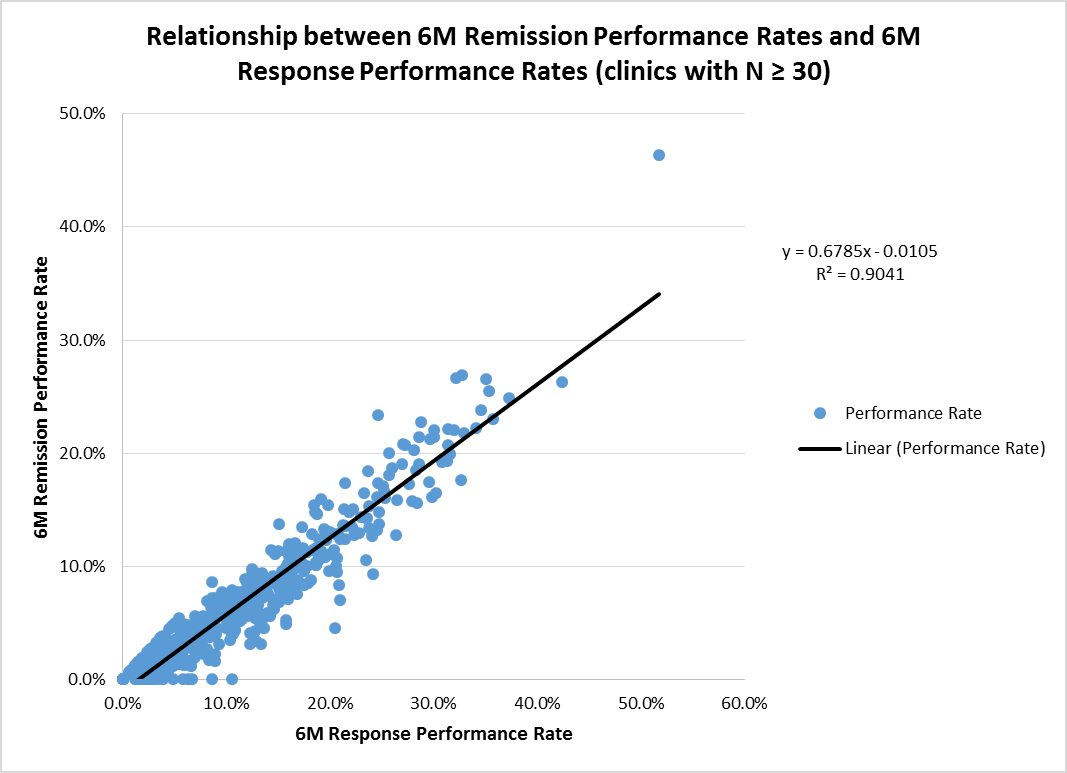
As PHQ-9 depression severity increased, there was a substantial decrease in functional status of all 6 SF-20 subscales in addition to an increase in symptom-related difficulty, sick days and health care utilization. Construct validity, using mental health professional re-interview as the criterion standard, 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 Performance Score

Correlation between Similar Depression Measures (Depression Response at 6 Months)



Validity testing of critical data elements:

For the quality measure, MNCM validates submitted data are complete via a four-step process that includes 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-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.

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.

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).

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 stratification)

* Strong correlation (R2 = 0.904) between clinics that do well with response rates (PHQ-9 initial score is reduced by 50% or greater) also are more successful at achieving remission.
* 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 stratification):

High compliance with critical data element validity as demonstrated by annual validation audit processes. High validity performance scores demonstrated through correlation (R2 = 0.904) with similar depression outcome measure.

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*)

There are two methods for excluding patients from the denominator for this longitudinal measure, upfront when the condition is known or along with the contact level submission when a condition becomes apparent during the index and assessment period. 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. When exclusions occur after the index contact event, they are included in the data submission for this measure and are available for analysis. In 2014, 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. Data is not yet available for re-designed exclusions to be implemented in report year 2020 (schizophrenia or psychotic disorder and pervasive developmental disorder)

Results of Testing Exclusions in Adolescent Population:

Of the 786 patients who met the denominator criteria across the three testing sites, 102 (13.0%) met the exclusion criteria of having a diagnosis of any of the following: bipolar disorder, psychotic disorder, autism spectrum disorder, or personality disorder. Of those patients excluded, 55 (53.9%) had a diagnosis of bipolar disorder.

Required exclusions include:

* Patient had an active diagnosis of bipolar disorder (*Bipolar Disorder* Value *Set*)
* Patient had an active diagnosis of schizophrenia or psychotic Disorder (*Schizophrenia or Psychotic Disorder* Value Set)

Allowable exclusions include:

* Patient was a permanent nursing home resident any time during the measurement period
* Patient was in hospice or receiving palliative care any time during the measurement period
* Patient died prior to the end of the measurement period
* Patient had an active diagnosis of personality disorder (*Personality Disorder* Value Set)
* Patient had an active diagnosis of Pervasive Developmental Disorder (Pervasive Disorder Value Set)

**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** three (severity of depression, age and insurance product) **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**.   
NA

**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*)

Note: Risk adjustment models related to the adult stratification and variables of severity of depression, age range and insurance products and are based on currently available data for the adult population stratification. Original documentation of testing for selection and application of risk adjustment variables remains and updated chi-square t-tests are provided for adults. It is anticipated that these variables will also work well for the adolescent stratification, and model testing will occur when adolescent data are available.

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 0.01.

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

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.

MNCM has incorporated risk adjustment for its publicly reported outcome measures using a staged approach. As other depression measures were included in the risk adjustment strategy, the variables for each measure were subject to significance testing with chi square t test < 0.01.

**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*)

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.

*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*)**:**Risk adjustment is necessary only when there is heterogeneity across clinics. There was significant heterogeneity across clinics in insurance product mix (χ2 = 10120, p < .001), patient

age (χ2 = 5325, p < .001), gender (χ2 = 1267, p < .001), initial severity (χ2 = 1759, p < .001), and distance to the clinic (χ2 = 35015, p < .001).

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Original Depression Remission at Six Months [2010]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 1: EFFECT OF POTENTIAL RISK ADJUSTERS ON DEPRESSION | | | | | | |
| **Variable** | **Contrast** | **Estimate** | **T-value** | **Odds Ratio** | **Lower 95% CI** | **Upper 95% CI** |
| 1A: MODEL WITHOUT SES AND RACE FROM ZIP CODE DATA | | | | | | |
| Age |  |  |  |  |  |  |
| 18-25 | 66+ | -0.62\*\* | -6.66 | 0.54\*\* | 0.45 | 0.65 |
| 26-50 | 66+ | -0.68\*\* | -8.70 | 0.51\*\* | 0.44 | 0.59 |
| 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 |
| Comorbidity |  |  |  |  |  |  |
| Moderate | Severe | 0.57\*\* | 10.56 | 1.77\*\* | 1.59 | 1.97 |
| 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 |
| 5-10 miles | Same Zip | -0.09 | -1.44 | 0.92 | 0.82 | 1.03 |
| 10-20 miles | Same Zip | -0.06 | -1.03 | 0.94 | 0.83 | 1.06 |
| 20+ miles | Same Zip | -0.10 | -1.33 | 0.90 | 0.78 | 1.05 |
| **Variable** | **Contrast** | **Estimate** | **T-value** | **Odds Ratio** | **Lower 95% CI** | **Upper 95% CI** |
| Insurance |  |  |  |  |  |  |
| Medicare | Commercial | -0.48\*\* | -9.72 | 0.55\*\* | 0.48 | 0.63 |
| Medicaid / MSHO / Special Needs / Self- pay / Uninsured | Commercial | -0.59\*\* | -8.83 | 0.62\*\* | 0.56 | 0.68 |
| Constant |  | -1.85 | -1.76 |  |  |  |
| 1B: MODEL WITH SES AND RACE FROM ZIP CODE DATA | | | | | | |
| Age |  |  |  |  |  |  |
| 18-25 | 66+ | -0.62\*\* | -6.65 |  |  |  |
| 26-50 | 66+ | -0.68\*\* | -8.67 |  |  |  |
| 51-65 | 66+ | -0.66\*\* | -8.47 |  |  |  |
| Gender |  |  |  |  |  |  |
| Female | Male | -0.08 | -1.87 |  |  |  |
| Comorbidity |  |  |  |  |  |  |
| Moderate | Severe | 0.57\*\* | 10.51 |  |  |  |
| Moderately Severe | Severe | 0.39\*\* | 6.80 |  |  |  |
| Distance from Clinic |  |  |  |  |  |  |
| <5 miles | Same Zip | -0.05 | -0.78 |  |  |  |
| 5-10 miles | Same Zip | -0.09 | -1.45 |  |  |  |
| 10-20 miles | Same Zip | -0.07 | -1.19 |  |  |  |
| 20+ miles | Same Zip | -0.09 | -1.22 |  |  |  |
| Insurance |  |  |  |  |  |  |
| Medicare | Commercial | -0.59\*\* | -8.74 |  |  |  |
| Medicaid / MSHO / Special Needs / Self- pay / Uninsured | Commercial | -0.46\*\* | -9.43 |  |  |  |
| Zip Code Data |  |  |  |  |  |  |
| Median Income |  | 0.03 | 1.91 |  |  |  |
| Percent Black |  | 0.00 | -0.81 |  |  |  |
| Percent Hispanic |  | 0.38 | 0.72 |  |  |  |
| Constant |  | -1.99 | -1.89 |  |  |  |
| \*\* indicates statistical significance. | | | | | | |

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Original Depression Remission at Six Months

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TABLE 2 – COMPARISON OF UNADJUSTED AND ADJUSTED DECILE RANKS (N / PERCENT OF ROW)\* | | | | | | | | | | | |
| Unadjusted Decile Rank | Risk Adjusted 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  66.67 | 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  3.57 | 0  0.00 | 0  0.00 | 0  0.00 | 28 |
| Unadjusted Decile Rank | Risk Adjusted 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 |
| 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  13.04 | 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 |
| \*Grey cells indicate no change in rank. Blue cells indicate increase in rank after risk adjustment, and red cells indicate decrease in rank after risk adjustment. N is the number of clinics in each cell, and the percent of row is the percent of the total unadjusted decile ranked clinics in each cell. | | | | | | | | | | | |

**Depression at 6 months Remission**

***Risk Variables: Product, Severity, Age Band***

Clinic Distribution 2013 Dates

**Result Compared to Expected**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **After Risk Adjustment** | | | |
| **Before RA** | Below | Expected | Above | Total |
| Significantly Below | 63 | 63 | 0 | 126 |
| Average | 0 | 234 | 0 | 234 |
| Significantly Above | 0 | 13 | 81 | 94 |
|  | 63 | 310 | 81 | 454 |
|  |  |  |  |  |
| Better | 63 | 14% |  |  |
| Same | 378 | 83% |  |  |
| Worse | 13 | 3% |  |  |

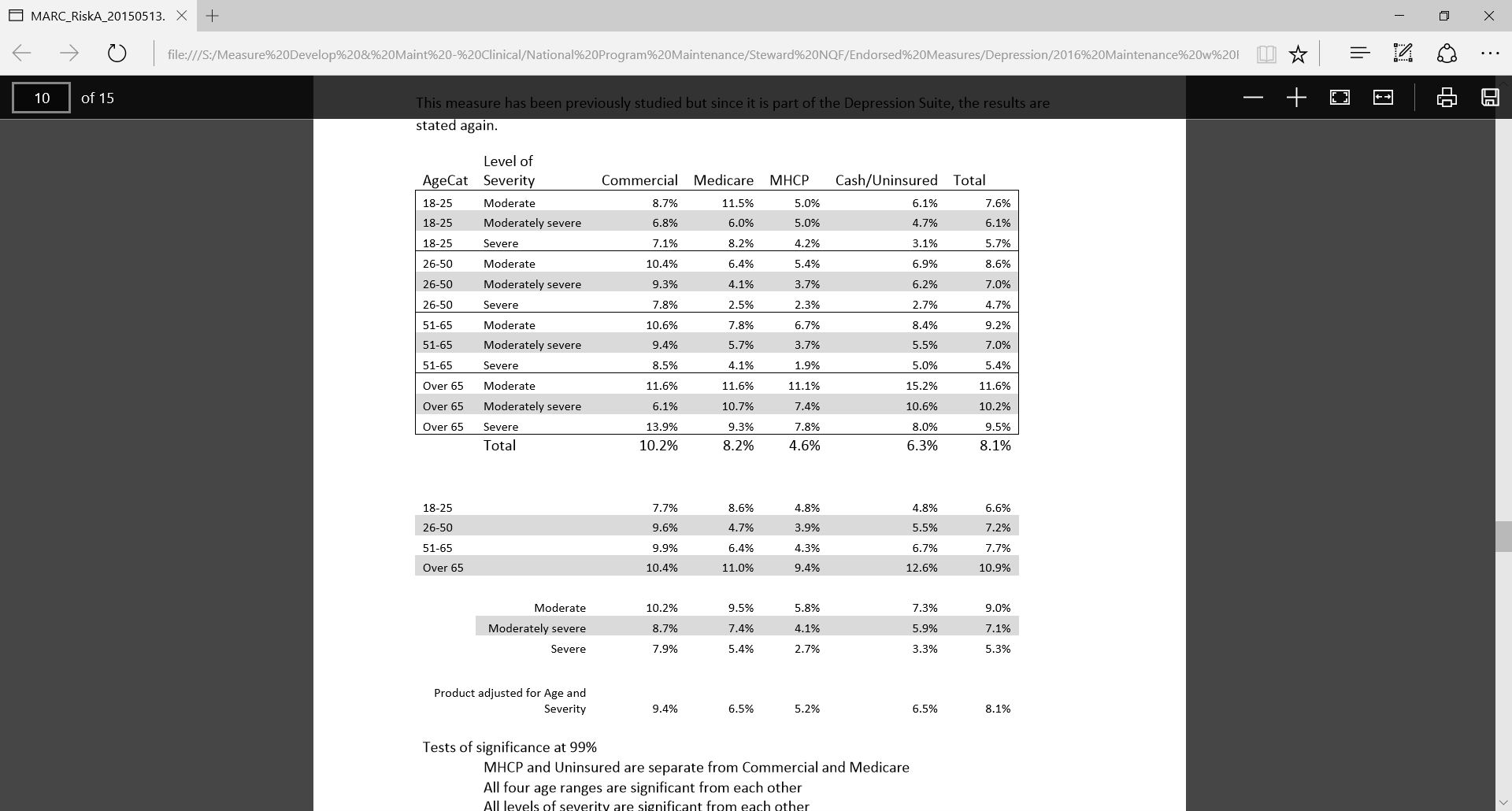
Updated Analysis [2014]

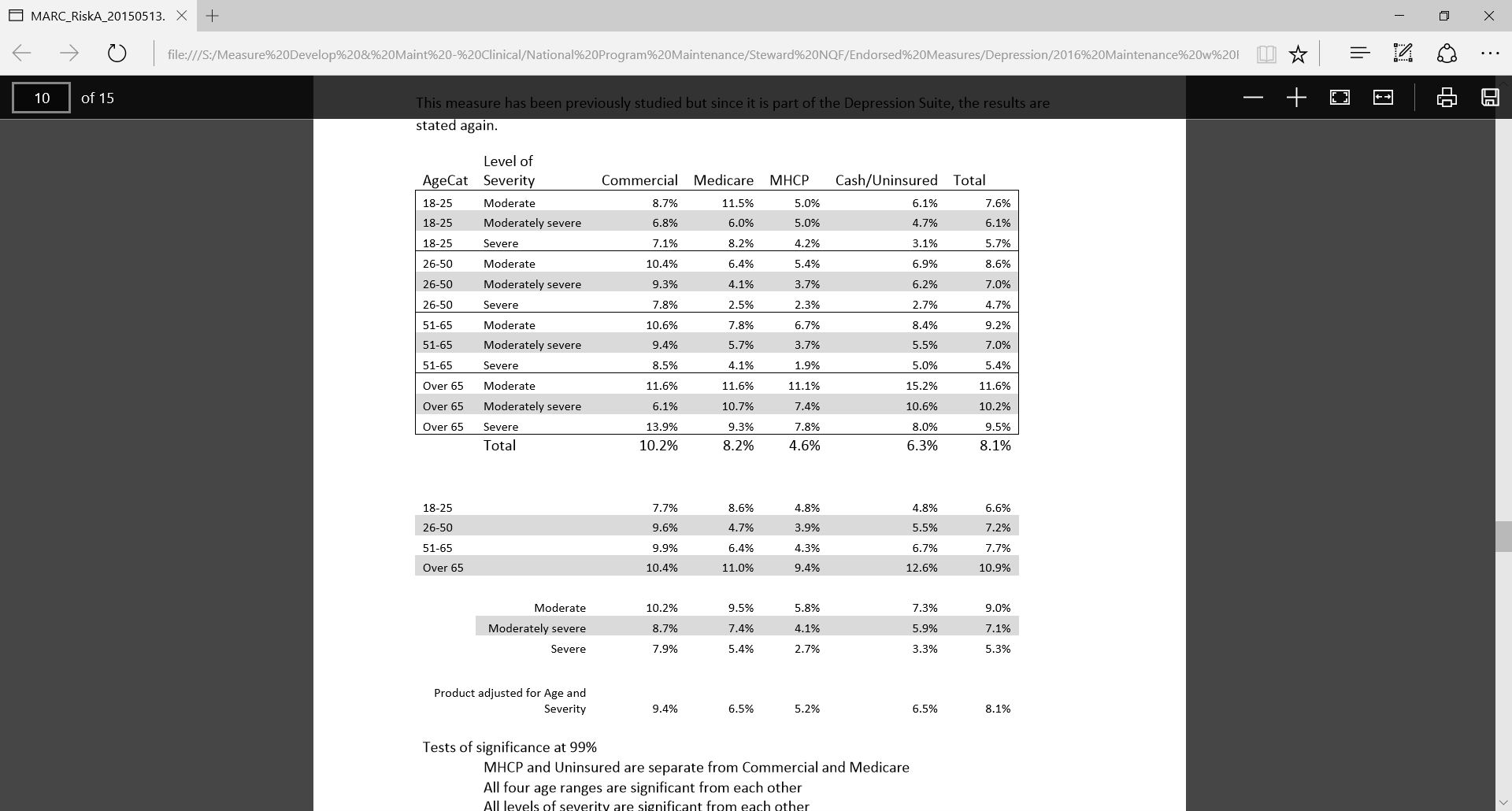
  
Older patients within the measure are starting with a lower level of depression. MHCP patients, even with adjusting for age, are starting with a higher level of depression. As you will see, severity does not have an impact on Follow Up or Response but does have a major impact on Remission.

There is a noticeably higher rate of Medicare patients under 65 for the depression measures than for any of the other MNCM Direct Data Submission measures.

Updated Analysis [2014]: Depression Remission at Six Months

Segmented by Risk Categories





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

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.

Updated [2014]: Depression Remission at Six Months

Summary of Variables Significance Testing (T test at .01)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  | |  |
|  | **6 Month Depression Measure Suite** | | | |
|  | **Follow Up** | **Response** | **Remission** | |
| **Product - Compared to Commercial** |  |  |  | |
| Medicare | <=.01 | 0.01 | <=.01 | |
| MHCP | <=.01 | <=.01 | <=.01 | |
| Uninsured | <=.01 | <=.01 | <=.01 | |
|  |  |  |  | |
| **Age - Compared to 51-65** |  |  |  | |
| 18-25 | <=.01 | <=.01 | <=.01 | |
| 26-50 | <=.01 | <=.01 | <=.01 | |
| Over 65 | 0.01 | <=.01 | <=.01 | |
|  |  |  |  | |
| **Severity- Compared to Moderate** |  |  |  | |
| Moderately Severe | 0.577 | 0.01 | <=.01 | |
| Severe | 0.04 | 0.541 | <=.01 | |

**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*)

Our analysis of risk adjustment factors for the Depression Remission at Six Months measure indicates that age, depression severity at diagnosis and some insurance provider variables (MSHO, Medicaid, and Medicare) are related to depression remission.

Tests of significance at .01%:

* MHCP and Uninsured are separate from Commercial and Medicare
* All four age ranges are significant from each other
* All levels of severity are significant from each other

Since the requirement for remission is for the follow up PHQ-9 to be below the specific level of 5 and not a percentage improvement, it makes sense that the initial level of severity is a significant factor in the outcome of remission. Analysis demonstrates a difference in severity of depression between populations; higher severity levels of depression are noted for MHCP and Uninsured patients. Age is also a significant factor that affects the outcome.

After analyzing the entire Depression suite of measures, it was reconfirmed that Age, Product and Severity Levels 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.

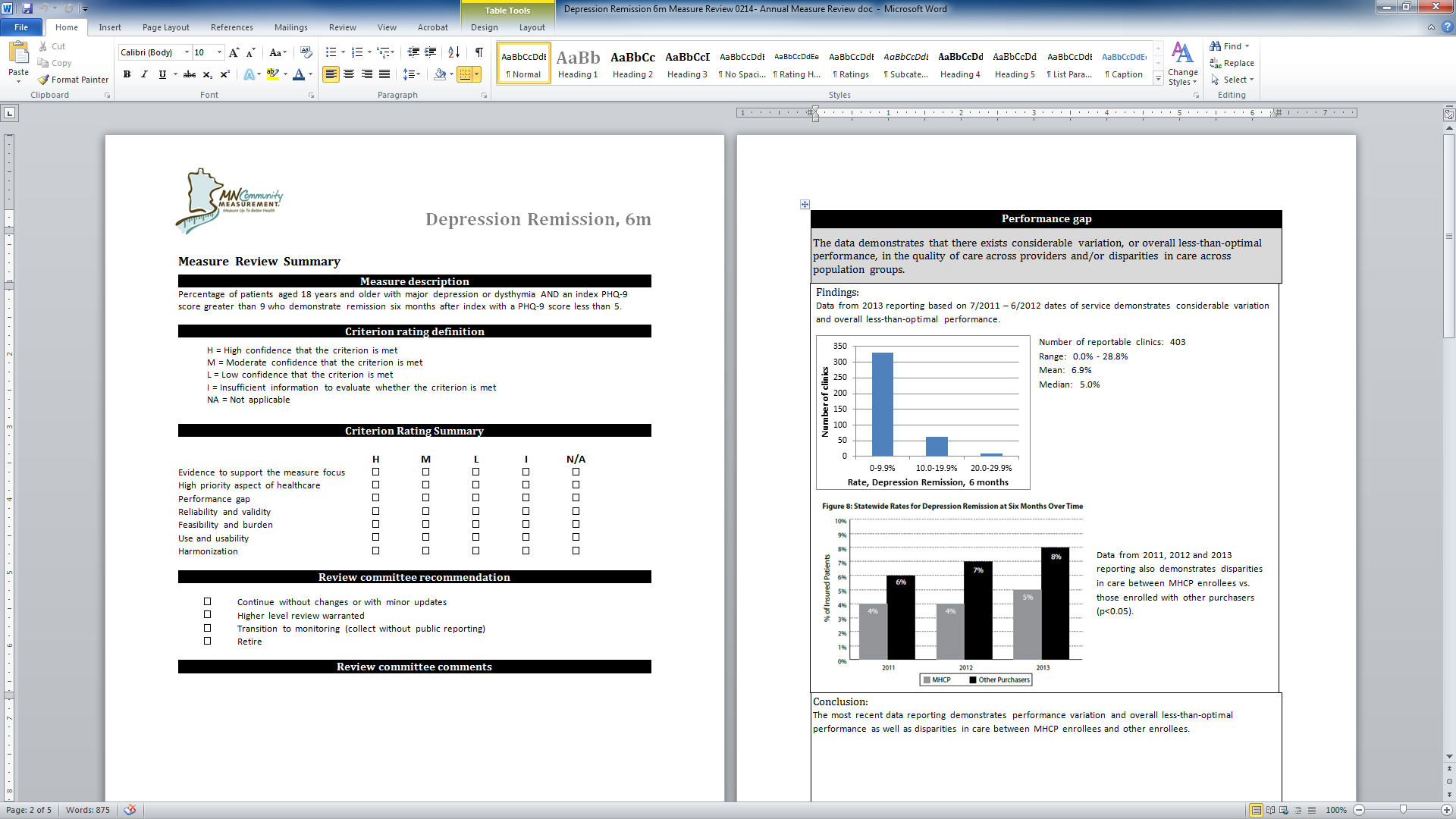
**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

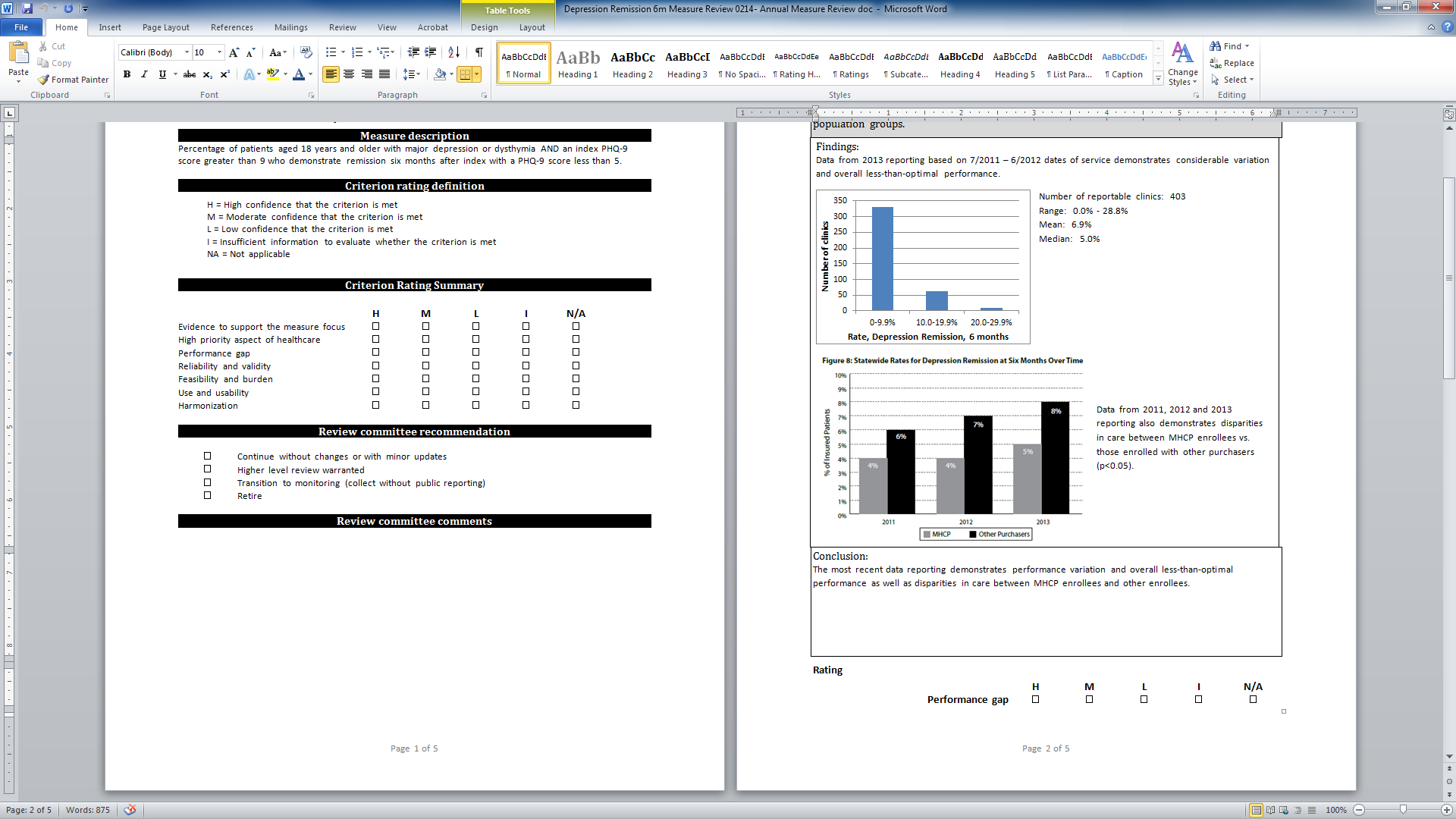
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

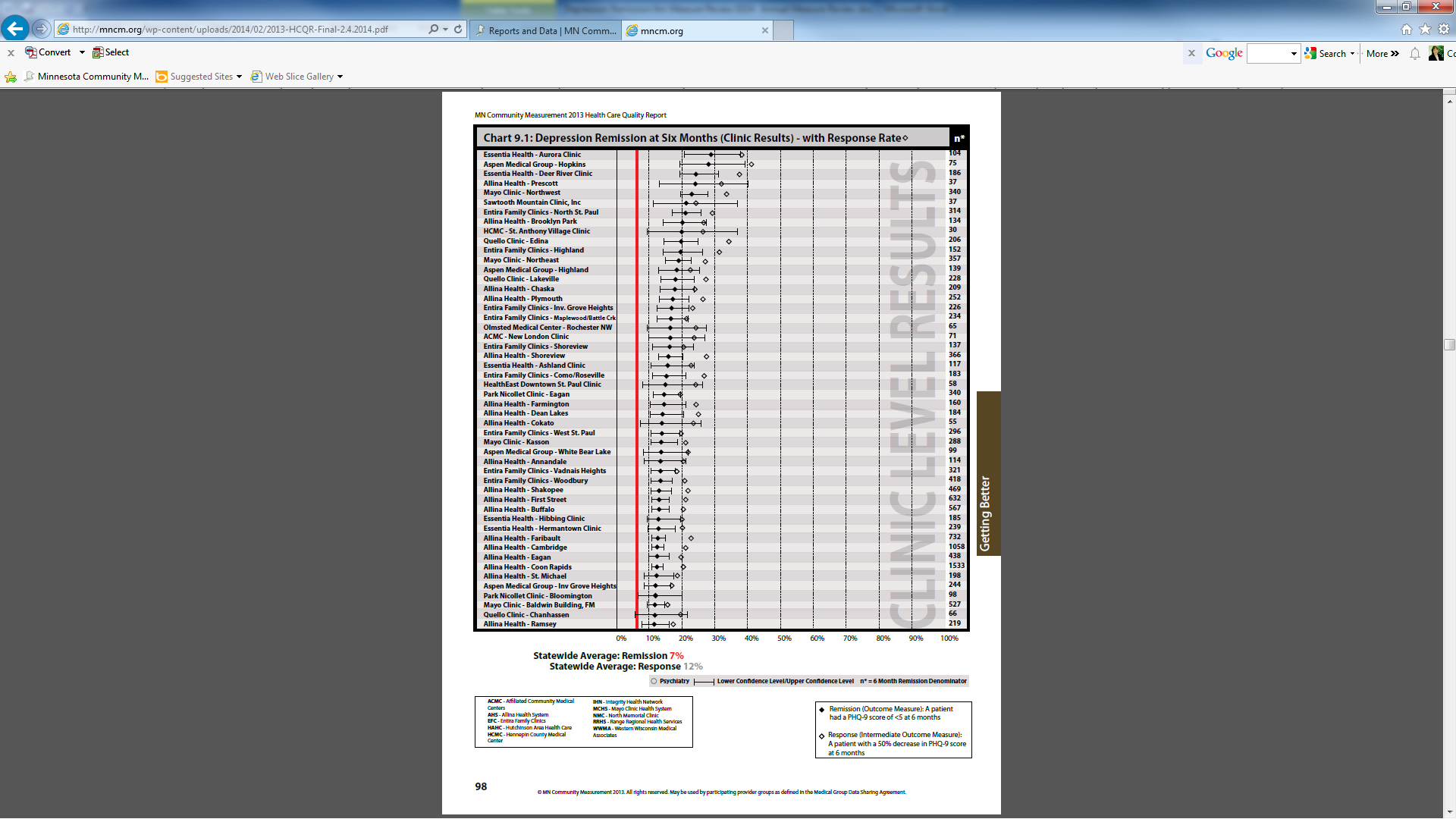
**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat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

Measure continues to demonstrate significant opportunity for both maintaining contact with the patients with depression (ongoing follow-up) and achieving an outcome of remission. Meaningful differences are demonstrated by:





Confidence intervals (95%) are displayed in our annual Health Care Quality Report at <http://mncm.org/wp-content/uploads/2014/02/2013-HCQR-Final-2.4.2014.pdf>. Additionally, this report highlights high performers, most improves and performance that has declined. Example below.



Identifying High Performing Medical Groups/Clinics

Medical groups that had rates that were fully above the medical group average and 95 percent confidence intervals were noted as high performers. These high performers were highlighted in the results by measure section for each measure. The performance rates for all medical groups were presented in descending order.

Identifying Medical Groups and Clinics with Biggest Improvements

For each measure, individual medical group and clinic rates during report year 2013 were compared with their rates during report year 2012, calculating an absolute percentage point difference. Medical groups and clinics with the largest percentage point increases were identified.

Medical Group and Clinic Performance Over Time (Three Years)

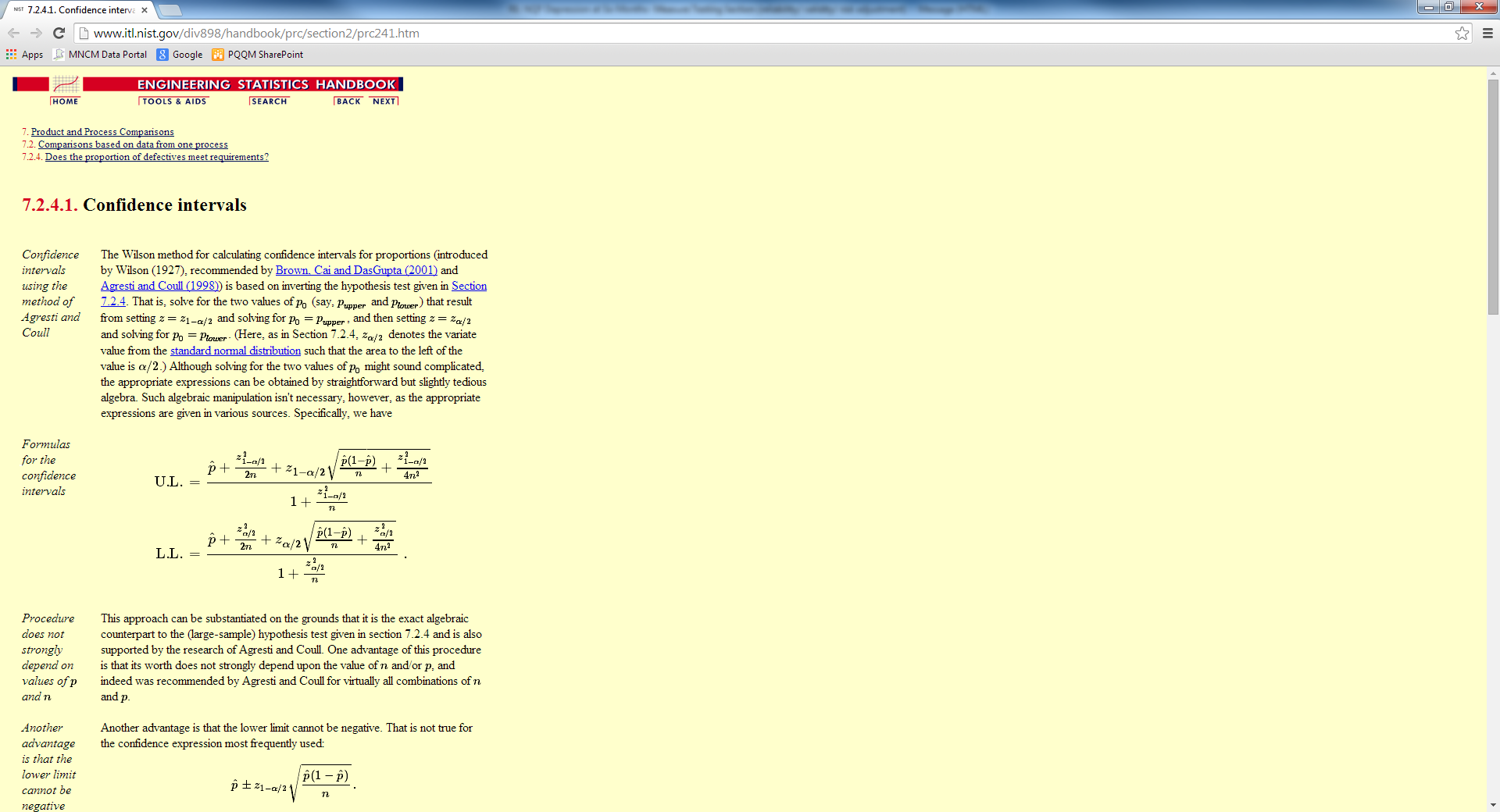
This analysis was done to determine patterns of medical group and clinic performance over time per measure. Patterns were reviewed for the three reporting years (2011, 2012 and 2013).

• Consistently improved: Medical groups with more than a two percentage point increase between each consecutive year.

• Relatively stable: Medical groups that had no more than a two percentage point increase or decrease between each consecutive year (-2 percent − +2 percent).

• Consistently decreased: Medical groups with more than a two percentage point decrease between each consecutive year.

• Variable performance (with an improvement or with a decline): Medical groups with an up/down pattern that was not consistent and did not fall into one of the other categories.



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

[www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm](http://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm)

**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 above 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 6 months, from 20.5% in 2010 to 31% in 2014.

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. Low outcome rates are not solely attributed to lack of follow-up. Additional analysis of denominator patients who do have a follow PHQ-9 score at six months (+/-30 days), only 25% of patients are in remission with a PHQ-9 < 5 and another 25% of patients demonstrate PHQ- 9 scores between 15 and 27 (major to severe symptoms)

**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 not assessed with a follow-up PHQ-9 at six months (+/- 30 days) are included in the denominator and treated as if they are not in remission.

**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