NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b2-2b7, 2d)

Measure Number (*if previously endorsed*)**: 0729**
Composite Measure Title: Optimal Diabetes Care
Date of Submission: 12/3/2014
Composite Construction:
☐Two or more individual performance measure scores combined into one score
☒ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)
☐ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- 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 composite measures, sections 1, 2a2, 2b2, 2b3, 2b5, and 2d must be completed.**
- **For composites with outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitions** (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), validity (2b2-2b6), and composites (2d) 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 25 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.

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] 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] 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]
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]

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; [14,15] 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] 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.

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2d1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and
2d2.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(*if not conducted or results not adequate, justification must be submitted and accepted)*

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.

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 different components in the composite, indicate the component 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*).

Existing data set was used. Patient level data was submitted from 118 medical groups representing 580 clinics in Minnesota and bordering communities. Data submission is an annual process; for 2013 dates of service reported in 2014, 230,818 patients were submitted for rate calculation and this represents 97.2% of all eligible diabetic patients in MN. Sampling is allowed for those clinics with paper records or those who have not had their EMR in place for one year prior to the measurement period; in MN there are very few clinics not yet on an EMR. For 2013 dates of service 91% submitted total population, 7% submitted a sample, and 2% submitted a mix of total and sample. Data submission for this measure is mandatory in the state of MN by 2008 health reform legislation and the MN Department of Health Statewide Quality and Reporting Measurement System.

Patients:	230,818
Medical Groups:	118
Individual Clinics:	580

Patient level data files are submitted by medical groups to a HIPAA secure data portal for rate calculation.
Types of fields included in the submission for 2013 dates of service include the following:

Clinic ID ▪ Patient ID ▪ Patient Date of Birth ▪ Patient Gender Zip Code, Primary Residence ▪ Race/Ethnicity1 ▪ Race/Ethnicity2 ▪ Race/Ethnicity3 ▪ Race/Ethnicity4 ▪ Race/Ethnicity5 ▪ Country of Origin Code ▪ Country of Origin "Other" Description ▪ Preferred Language Code ▪ Preferred Language "Other" Description ▪ Provider NPI ▪ Provider Specialty Code ▪ Insurance Coverage Code ▪ Insurance Coverage "Other" Description ▪ Insurance Plan Member ID ▪ Patient Has IVD? ▪ Patient Has Depression? ▪ Type 1 or Type 2 Diabetes? ▪ HbA1c Date ▪ HbA1c Value ▪ LDL Date ▪ LDL Value ▪ BP Date ▪ BP Systolic ▪ BP Diastolic ▪ Aspirin (ASA) Date ▪ Aspirin (ASA) Contraindication Date ▪ Tobacco Status Documentation Date ▪ Tobacco Status

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

Patients: 230,818
Medical Groups: 118
Individual Clinics with ≥ 30 eligible:
580 clinic sites and 229,806 patients
Individual clinics with < 30 eligible:
77 clinic sites with 1,012 patients

Includes all primary care and endocrinology clinics in MN; rates for public reporting and associated reliability and clinic level statistics only include those clinic sites with 30 or more eligible diabetic patients meeting denominator criteria; in 2013 this was 580 clinics representing over 229,000 patients. All patients submitted are used to calculate statewide averages, risk adjustment models and aggregate descriptive statistics.

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*)

Patients: 230,818
Medical Groups: 118
Individual Clinics with ≥ 30 eligible:

580 clinic sites and 229,806 patients
Individual clinics with < 30 eligible:
77 clinic sites with 1,012 patients

Includes all primary care and endocrinology clinics in MN; rates for public reporting and associated reliability and clinic level statistics only include those clinic sites with 30 or more eligible diabetic patients meeting denominator criteria; in 2013 this was 580 clinics representing over 229,000 patients. All patients submitted are used to calculate statewide averages, risk adjustment models and aggregate descriptive statistics.

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.

_____

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted?

Note*: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. 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*)

Used paper "Reliability in Provider Profiling" by John L. Adams, Ph.D as a reference
The BETABIN macro was used on each measure (SAS).

- First, we need to find the provider-to-provider variance:
  - $\sigma^2 = (\alpha\ \beta) / (\alpha + \beta + 1)(\alpha + \beta)^2$
  - $= (6.3684 * 11.1559) / (6.3684 + 11.1559 + 1)(6.3684 + 11.1559)^2$
  - = **0.0125** (plug this value into the reliability equation)
- Reliability = $\sigma^2 / (\sigma^2 + (p(1 - p)/n))$
  - p = rate
  - n = number of eligible patients
- Determine reliability rate for each provider.
- Average the reliability rate.

2a2.3. 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*)

Reliability = 0.908

BETABIN Macro: Simple Binomial Model

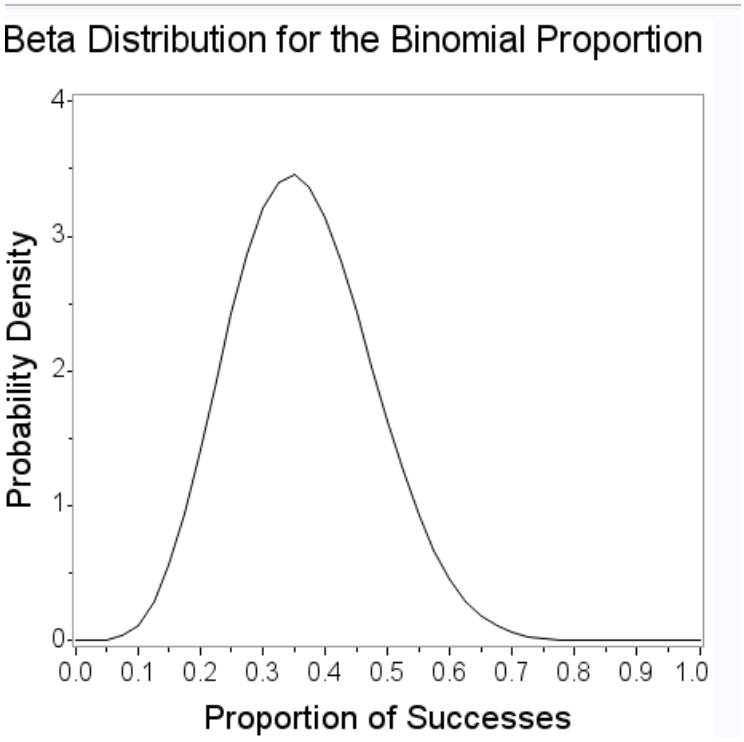
parameter	Estimate	Standard Error	t Value	Prob > |t|	Alpha	Lower	Upper
mu	0.3925	0.0010	385.34	<.0001	0.05	0.3905	0.3945
|mu-0.5|	0.1075	0.0010	105.51	<.0001	0.05	0.1055	0.1095

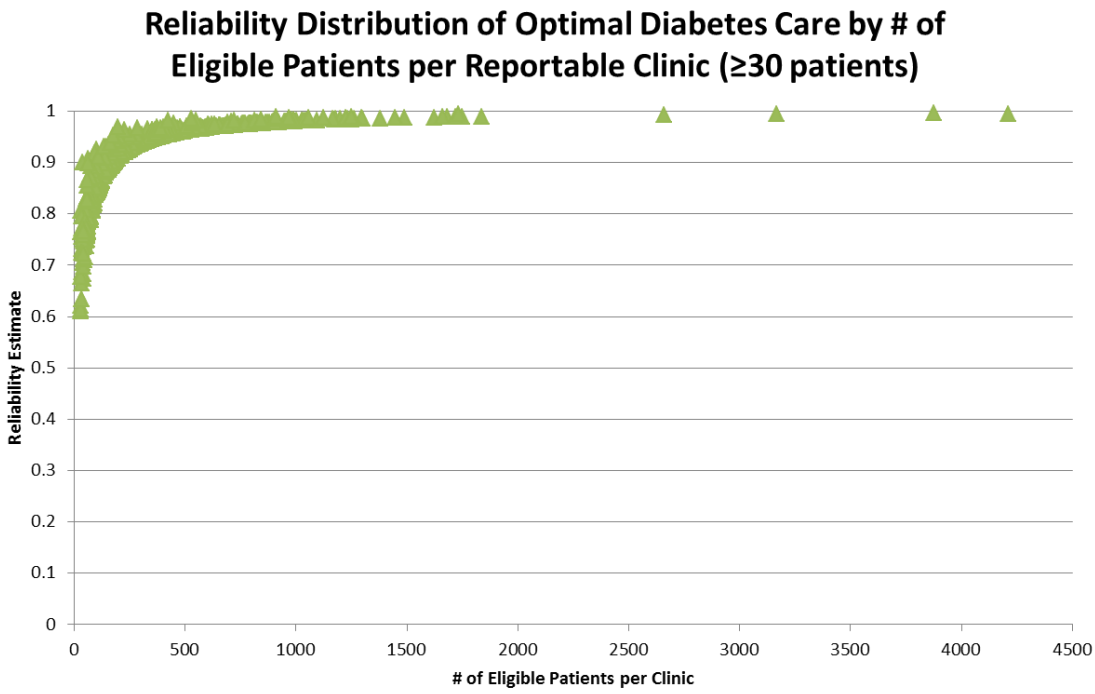
BETABIN Macro: Beta-Binomial Model Parameters

parameter	Estimate	Standard Error	t Value	Pr > |t|	Alpha	Lower	Upper
mu	0.3634	0.004878	74.50	<.0001	0.05	0.3538	0.3730
alpha	6.3684	0.4155	15.33	<.0001	0.05	5.5540	7.1828
beta	11.1559	0.7266	15.35	<.0001	0.05	9.7317	12.5800
gamma	0.05398	0.003288	16.42	<.0001	0.05	0.04754	0.06043
theta	0.05706	0.003674	15.53	<.0001	0.05	0.04986	0.06426
|mu-0.5|	0.1366	0.004878	28.00	<.0001	0.05	0.1270	0.1462

BETABIN Macro: Variance-Covariance Matrix of Estimated Parameters

Label	mu	alpha	beta	gamma	theta
mu	0.000024	0.000338	-0.00055	6.298E-7	7.037E-7
alpha	0.000338	0.1727	0.2861	-0.00134	-0.00149
beta	-0.00055	0.2861	0.5280	-0.00237	-0.00265
gamma	6.298E-7	-0.00134	-0.00237	0.000011	0.000012
theta	7.037E-7	-0.00149	-0.00265	0.000012	0.000013





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?*)

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.

2b2. VALIDITY TESTING

Note*: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b2.1. What level of validity testing was conducted?

☒ **Composite 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*)

☐ **Systematic assessment of content validity**

☒ **Validity testing for component measures** *(check all that apply)*

Note*: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☐ **Empirical validity testing of the component measure score(s)**
☐ **Systematic assessment of face validity of component measure score(s) 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)*

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
- Eligible specialties and provider types
- Exclusions to the measure and attest to mechanism for exclusions
- Attestations related to changes in medical record or billing systems
- Supplying all query code for review

Common areas of correction in denominator for this measure included missing query code, incorrect date of birth ranges, incorrect dates for counting visits, missing ICD-9 codes 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 diabetes measure, MNCM audited 128 medical groups; 76% of those submitting data. 85% passed the initial audit, 15% required a correction plan and all re-submitted their data and passed the audit with $\geq$ 90% accuracy. Types of discrepancies noted on audit included: not including most recent values, aspirin/ anti-platelet dates not in the measurement year and incorrect tobacco status. The error rate with performance score impact was 0.3%; however, this did not result in any medical group level score errors.

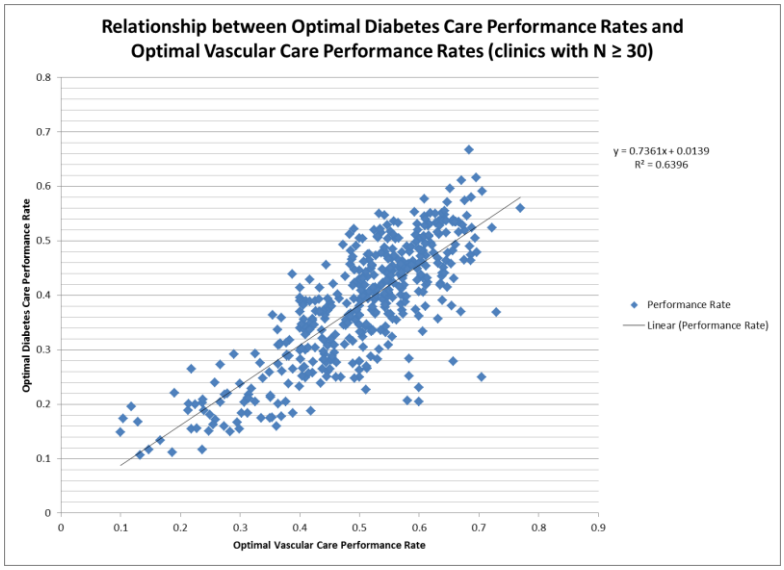
Validity was tested for the computed composite score by testing the correlation of medical group performance with their performance on the Optimal Vascular Care measure (NQF#0076). Ischemic vascular disease and diabetes are chronic conditions that require ongoing management of multiple risk

factors in order to reduce a patient's overall risk of developing long term complications. It is expected that the quality of care provided by a medical group to patient with diabetes would be of similar quality as the care provided to patients with ischemic vascular disease, and the respective performance measure scores should demonstrate such.

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

100% of groups achieved the desired > 90% data accuracy when submitted data was compared to medical record data (EMR or paper) of the patient.

Based on linear regression analysis, a medical group's performance on the Optimal Diabetes Care measure is associated with its performance on the Optimal Vascular Care measure, as demonstrated by an r^2 value of 64%, representing a fairly strong correlation.



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?*)

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

As demonstrated by the r^2 value, 64% of the total variation in performance on the Optimal Diabetes Care measure can be explained by variation in the Optimal Vascular Care measure. This high degree of correlation indicates that the Optimal Diabetes Care composite measure score accurately reflects the quality of care provided.

2b3. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

NA ☐ no exclusions — *skip to section 2b4*

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*)

Medical groups submitting patient level data to MNCM have the option of either excluding patients appropriately via their EMR query process and provide an attestation through our process of denominator certification and providing their query code for MNCM staff to review and/or submitting a file of excluded patients. Exclusion testing was performed on a sample of groups who submitted files of patients they excluded from the measure. Sample included 11 medical groups representing metro, rural, endocrinology, federally qualified health care centers and teaching/ tertiary. Included were 232 clinics and over 109,000 diabetic patients with dates of service 1/1/2013 to 12/31/2013.

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*)

Medical Group Code	Clinics	Type	Diabetic Patients	Nursing Home	Hospice	Deceased	Coded in error	Pregnancy	Total Exclusions
Medical Group A	51	Metro	32,110	0	0	266	0	118	384
Medical Group B	8	Rural	2,588	2	2	33	0	1	38
Medical Group C	12	Rural	3,757	24	1	52	0	8	85
Medical Group D	1	Metro; Endo	80	0	0	1	0	0	1
Medical Group E	39	Metro	17,289	0	0	124	0	30	154
Medical Group F	28	Metro	15,182	61	64	3	0	51	179
Medical Group G	13	Metro; FQHC	5,461	3	1	0	0	1	5
Medical Group H	12	Rural	2,490	0	0	20	0	1	21
Medical Group I	23	Metro	17,628	6	6	95	2	88	197
Medical Group J	44	Metro/ Rural	9,733	34	0	1	0	17	52
Medical Group K	1	Metro/ Teaching	3,346	0	0	1	7	8	16
	232		**109,664**	**130**	**74**	**596**	**9**	**323**	**1,132**
				0.1%	**0.1%**	**0.5%**	**0.01%**	**0.3%**	**1.0%**

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*)

While the exclusions to this measure are have clinical importance related to patient safety in achieving targets or utilizing medications to reduce cardiovascular risk, the total number of exclusions is relatively small, 1.0% and therefore do not significantly impact measure performance.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☐ **Endorsed (or submitted) as individual performance measures**
☐ **No risk adjustment or stratification**
☒ **Statistical risk model**
☐ **Stratification by risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

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*)
Normally, during our measure development process, the expert panel discusses potential variables for risk adjustment that are important to consider for the measured population. Variables are included in public comment and collected during pilot testing to assess feasibility. For this measure, which has been in place since ~ 2004, MNCM was not the developer but has taken on the stewardship. Variables for risk adjustment have been selected and tested over the years. In addition to our standard demographic variables (gender, age, zip, race/ethnicity, country of origin, primary language and insurance product), we have also assessed clinical variables including type of diabetes, major depression, and ischemic vascular disease. The potential risk adjustment variables are then evaluated for appropriate inclusion in the model based on a t value outside the range of -2.0 and +2.0.

The variables that have been selected are insurance product (commercial, medicare, MN government programs, self-pay/uninsured), age bands (18-25, 26-50, 51-65 and 65 to 75) and diabetes type (1 or 2)

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

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(11, 12). These two methods are compared directly in our accompanying report on the quality deviations ranking approach.

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*)

Because ODC is a binary variable (0 or 1), the risk adjustment model was estimated using a logistic model implemented in the SAS Procedure Glimmix that accounts for its non-continuous nature. The

risk adjusters and an indicator for each clinic were included in the model. The estimated coefficient for the clinic indicator measures the clinic's ODC adjusting for the patient risk adjusters that were included in the model. The clinic level indicator was used to construct a risk adjusted ODC score at the clinic level that ranged from 0 to 1 (0% to 100%). The effect of risk adjustment on clinic rankings was calculated by comparing the risk adjusted ODC to the unadjusted ODC measure, the average ODC for all patients reported by the clinic. The risk adjustment for tables 2-4 includes all risk adjustment variables detailed in Table 1 (age, gender, comorbidity, distance, and insurance). Since age greater than 65, the contrast category for age effects, captures the effect of age, the Medicare indicator for insurance captures the effect of Medicare independent of age.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
At the clinic level, the average ODC measure was 35.1% (standard deviation = 12%). The average number of patients reported by a clinic was 348 (standard deviation = 405). At the patient level, the average ODC was 39.7%. The average age in the examined population was 58, 46% were female, 7.3% had Type I diabetes, 19.6% were depressed, 48.8% had commercial insurance, 33.1% had Medicare coverage, and 7.1% had Medicaid coverage.

Risk adjustment is necessary only when there is heterogeneity across clinics. There was significant heterogeneity across clinics in insurance product mix ($\chi 2 = 65617$, $p < .001$), patient age ($\chi 2 = 12522$, $p < .001$), gender ($\chi 2 = 5256$, $p < .001$), depression ($\chi 2 = 4290$, $p < .001$), Type 1 Diabetes ($\chi 2 = 67297$, $p < .001$), and distance to the clinic ($\chi 2 = 63638$, $p < .001$).

We tested the overall correlation between the unadjusted and risk adjusted ODC 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 ODC values, and is .96 which shows a very strong correlation between the unadjusted and adjusted ODC measure. The Kendall's Tau correlation compares unadjusted and adjusted rank order of clinics, and was .85. This is still a strong correlation, but not as strong as the .96 correlation between risk adjusted and unadjusted clinic values.

We used various methods to compare the effect of risk adjustment on clinic rank (risk adjustment includes all variables detailed in Table 1), as shown in Tables 2 through 4. Table 2 compares the unadjusted and adjusted decile ranking of clinics (the decile approach). Table 3 compares unadjusted and adjusted clinic quality rankings based on their statistical difference from the ODC population mean (the quality deviations approach). Table 4 compares the adjusted clinic rankings between both the decile and quality deviations approaches.

Our analysis of the decile approach shows that, consistent with the Kendall's Tau correlation analysis, there are not major differences between the adjusted and unadjusted clinic rankings by decile (shown in Table 2). Most clinics (309) remain in the diagonal, which indicates no change in clinic ranking due to risk adjustment, while some (101) increase in ranking and others (116) decrease in ranking. Table 3 compares the unadjusted and adjusted clinic rankings using the quality deviations approach. Consistent with the decile ranking approach, Table 3 shows that the majority of clinics (410) experience no change in clinic ranking due to risk adjustment while a few (58) increase in rank and a few (58) decrease in rank.

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

TABLE 1: EFFECT OF POTENTIAL RISK ADJUSTERS ON ODC

1A: MODEL WITHOUT SES AND RACE FROM ZIP CODE DATA

Variable	Contrast	Estimate	T-value	Odds Ratio	Lower 95% CI	Upper 95% CI
Age						
18-25	66-75	-1.46**	-21.98	0.23**	0.20	0.26
26-50	66-75	-0.91**	-50.75	0.40**	0.39	0.42
51-65	66-75	-0.44**	-30.22	0.64**	0.63	0.66
Gender						
Female	Male	-0.04**	-3.68	0.96**	0.94	0.98
Comorbidity						
Depressed	Not Depressed	-0.21**	-15.41	0.81**	0.79	0.83
Type 1 Diabetes	Type 2 Diabetes	-0.29**	-12.54	0.75**	0.71	0.78
Distance from Clinic						
<5 miles	Same Zip	0.01	0.31	1.01	0.97	1.04
5-10 miles	Same Zip	0.00	-0.21	1.00	0.97	1.03
10-20 miles	Same Zip	-0.03*	-2.02	0.97	0.94	1.00
20+ miles	Same Zip	-0.10**	-5.00	0.90**	0.87	0.94
Insurance						
Medicare	Commercial	-0.06**	-4.26	0.94**	0.91	0.97
Medicaid / MSHO / Special Needs / Self-pay / Uninsured	Commercial	-0.52**	-31.07	0.60**	0.58	0.62
Constant		-0.39	-1.33			

1B: MODEL WITH SES AND RACE FROM ZIP CODE DATA

Age			
18-25	66-75	-1.45**	-21.85
26-50	66-75	-0.90**	-50.22
51-65	66-75	-0.44**	-29.91
Gender			
Female	Male	-0.04**	-3.41
Comorbidity			
Depressed	Not Depressed	-0.21**	-15.28
Type 1 Diabetes	Type 2 Diabetes	-0.30**	-12.74
Distance from Clinic			
<5 miles	Same Zip	0.01	0.70
5-10 miles	Same Zip	-0.02	-0.95
10-20 miles	Same Zip	-0.06**	-3.35
20+ miles	Same Zip	-0.10**	-4.98
Insurance			
Medicare	Commercial	-0.05**	-3.76
Medicaid / MSHO / Special Needs / Self-pay / Uninsured	Commercial	-0.50**	-30.04
Zip Code Data			
Median Income		0,03**	6.70
Percent Black		-.01**	-4.96
Percent Hispanic		0.06	0.41
Constant		-0.53	-1.79

** indicates statistical significance.

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

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%	40 76.92	10 19.23	2 3.85	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	52
10% to 20%	12 21.43	34 60.71	7 12.50	3 5.36	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	56
20% to 30%	0 0.00	9 18.37	32 65.31	7 14.29	1 2.04	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	49
30% to 40%	0 0.00	0 0.00	11 20.75	30 56.60	11 20.75	1 1.89	0 0.00	0 0.00	0 0.00	0 0.00	53
40% to 50%	0 0.00	0 0.00	1 1.89	11 20.75	24 45.28	12 22.64	3 5.66	1 1.89	1 1.89	0 0.00	53
50% to 60%	0 0.00	0 0.00	0 0.00	1 1.92	16 30.77	26 50.00	6 11.54	1 1.92	2 3.85	0 0.00	52
60% to 70%	0 0.00	0 0.00	0 0.00	0 0.00	1 1.89	13 24.53	28 52.83	9 16.98	2 3.77	0 0.00	53
70% to 80%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	15 28.30	28 52.83	10 18.87	0 0.00	53
80% to 90%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 1.89	0 0.00	14 26.42	28 52.83	10 18.87	53
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	10 19.23	42 80.77	52
Total	52	53	53	52	53	53	52	53	53	52	526

*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.

RISK ADJUSTMENT FOR OPTIMAL DIABETES CARE QUALITY MEASURE

TABLE 3 – COMPARISON OF UNADJUSTED AND ADJUSTED DEVIATIONS QUALITY RANKINGS (N / PERCENT OF ROW)

Unadjusted Deviations Quality	Risk Adjusted Deviations Quality					
	Poor (3+ SD below mean)	Below Average (2-3 SD below mean)	Average (mean + or – 2 SD)	Above Average (2-3 SD above mean)	Excellent (3+ SD above mean)	Total
Poor (3+ SD below mean)	50 64.10	27 34.62	1 1.28	0 0.00	0 0.00	78
Below Average (2-3 SD below mean)	1 1.96	37 72.55	13 25.49	0 0.00	0 0.00	51
Average (mean + or – 2 SD)	1 0.50	7 3.48	186 92.54	5 2.49	2 1.00	201
Above Average (2-3 SD above mean)	0 0.00	0 0.00	26 60.47	13 30.23	4 9.30	43
Excellent (3+ SD above mean)	0 0.00	0 0.00	2 1.31	18 11.76	133 86.93	153
Total	52	71	228	36	139	526

* 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.

TABLE 4 – COMPARISON OF RISK ADJUSTED DECILE AND DEVIATIONS QUALITY RANKINGS (N / PERCENT OF ROW)

Risk Adjusted Decile	Risk Adjusted Deviations Quality					
	Poor (3+ SD below mean)	Below Average (2-3 SD below mean)	Average (mean + or – 2 SD)	Above Average (2-3 SD above mean)	Excellent (3+ SD above mean)	Total
0 to 10%	28 53.85	23 44.23	1 1.92	0 0.00	0 0.00	52
10% to 20%	10 18.87	28 52.83	15 28.30	0 0.00	0 0.00	53
20% to 30%	11 20.75	8 15.09	34 64.15	0 0.00	0 0.00	53
30% to 40%	2 3.85	12 23.08	38 73.08	0 0.00	0 0.00	52
40% to 50%	1 1.89	0 0.00	52 98.11	0 0.00	0 0.00	53
50% to 60%	0 0.00	0 0.00	52 98.11	1 1.89	0 0.00	53
60% to 70%	0 0.00	0 0.00	24 46.15	16 30.77	12 23.08	52
70% to 80%	0 0.00	0 0.00	8 15.09	9 16.98	36 67.92	53
80% to 90%	0 0.00	0 0.00	4 7.55	6 11.32	43 81.13	53
90% to 100%	0 0.00	0 0.00	0 0.00	4 7.69	48 92.31	52
Total	52	71	228	36	139	526

*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.

RISK ADJUSTMENT FOR OPTIMAL DIABETES CARE QUALITY MEASURE

TABLE 5 – COMPARISON OF UNADJUSTED AND PRODUCT 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%	41 78.85	11 21.15	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	52
10% to 20%	11 19.64	34 60.71	10 17.86	1 1.79	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	56
20% to 30%	0 0.00	8 16.33	34 69.39	7 14.29	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	49
30% to 40%	0 0.00	0 0.00	9 16.98	32 60.38	12 22.64	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	53
40% to 50%	0 0.00	0 0.00	0 0.00	12 22.64	27 50.94	10 18.87	2 3.77	2 3.77	0 0.00	0 0.00	53
50% to 60%	0 0.00	0 0.00	0 0.00	0 0.00	14 26.92	32 61.54	2 3.85	2 3.85	1 1.92	1 1.92	52
60% to 70%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	11 20.75	38 71.70	3 5.66	1 1.89	0 0.00	53
70% to 80%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	9 16.98	37 69.81	7 13.21	0 0.00	53
80% to 90%	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 1.89	9 16.98	38 71.70	5 9.43	53
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	6 11.54	46 88.46	52
Total	52	53	53	52	53	53	52	53	53	52	526

*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.

TABLE 6 – COMPARISON OF UNADJUSTED AND PRODUCT ADJUSTED DEVIATIONS QUALITY RANKINGS (N / PERCENT OF ROW)

Unadjusted Deviations Quality	Risk Adjusted Deviations Quality					
	Poor (3+ SD below mean)	Below Average (2-3 SD below mean)	Average (mean + or – 2 SD)	Above Average (2-3 SD above mean)	Excellent (3+ SD above mean)	Total
Poor (3+ SD below mean)	68 87.18	10 12.82	0 0.00	0 0.00	0 0.00	78
Below Average (2-3 SD below mean)	5 9.80	40 78.43	6 11.76	0 0.00	0 0.00	51
Average (mean + or – 2 SD)	1 0.50	8 3.98	186 92.54	4 1.99	2 1.00	201
Above Average (2-3 SD above mean)	0 0.00	0 0.00	10 23.26	31 72.09	2 4.65	43
Excellent (3+ SD above mean)	0 0.00	0 0.00	0 0.00	7 4.58	146 95.42	153
Total	74	58	202	42	150	526

* 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.

2b4.9. Results of Risk Stratification Analysis:

At the clinic level, the average ODC measure was 35.1% (standard deviation = 12%). The average number of patients reported by a clinic was 348 (standard deviation = 405). At the patient level, the average ODC was 39.7%. The average age in the examined population was 58, 46% were female, 7.3% had Type I diabetes, 19.6% were depressed, 48.8% had commercial insurance, 33.1% had Medicare coverage, and 7.1% had Medicaid coverage.

Risk adjustment is necessary only when there is heterogeneity across clinics. There was significant heterogeneity across clinics in insurance product mix ($\chi2 = 65617$, $p < .001$), patient age ($\chi2 = 12522$, $p < .001$), gender ($\chi2 = 5256$, $p < .001$), depression ($\chi2 = 4290$, $p < .001$), Type 1 Diabetes ($\chi2 = 67297$, $p < .001$), and distance to the clinic ($\chi2 = 63638$, $p < .001$).

We tested the overall correlation between the unadjusted and risk adjusted ODC 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 ODC values, and is .96 which shows a very strong correlation between the unadjusted and adjusted ODC measure. The Kendall's Tau correlation compares unadjusted and adjusted rank order of clinics, and was .85. This is still a strong correlation, but not as strong as the .96 correlation between risk adjusted and unadjusted clinic values.

We used various methods to compare the effect of risk adjustment on clinic rank (risk adjustment includes all variables detailed in Table 1), as shown in Tables 2 through 4. Table 2 compares the unadjusted and adjusted decile ranking of clinics (the decile approach). Table 3 compares unadjusted and adjusted clinic quality rankings based on their statistical difference from the ODC population mean (the quality deviations approach). Table 4 compares the adjusted clinic rankings between both

the decile and quality deviations approaches.

Our analysis of the decile approach shows that, consistent with the Kendall's Tau correlation analysis, there are not major differences between the adjusted and unadjusted clinic rankings by decile (shown in Table 2). Most clinics (309) remain in the diagonal, which indicates no change in clinic ranking due to risk adjustment, while some (101) increase in ranking and others (116) decrease in ranking. Table 3 compares the unadjusted and adjusted clinic rankings using the quality deviations approach. Consistent with the decile ranking approach, Table 3 shows that the majority of clinics (410) experience no change in clinic ranking due to risk adjustment while a few (58) increase in rank and a few (58) decrease in rank.

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 Optimal Diabetes Care measure indicates that age, gender, comorbidity, distance of patient residence from clinic, and insurance provider variables are related to ODC and may warrant attention for risk adjustment.

For most clinics there is no change in clinic ranking due to risk adjustment, while some increase in ranking and others decrease in ranking. For those whose rankings are impacted by the risk adjustment, it is legitimate and based on disparate differences among these clinics.

***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
Note: *Applies to the composite performance measure.*

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 opportunity for improvement as well as statistically significant and clinically meaningful differences between medical group practices and clinics.

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*)

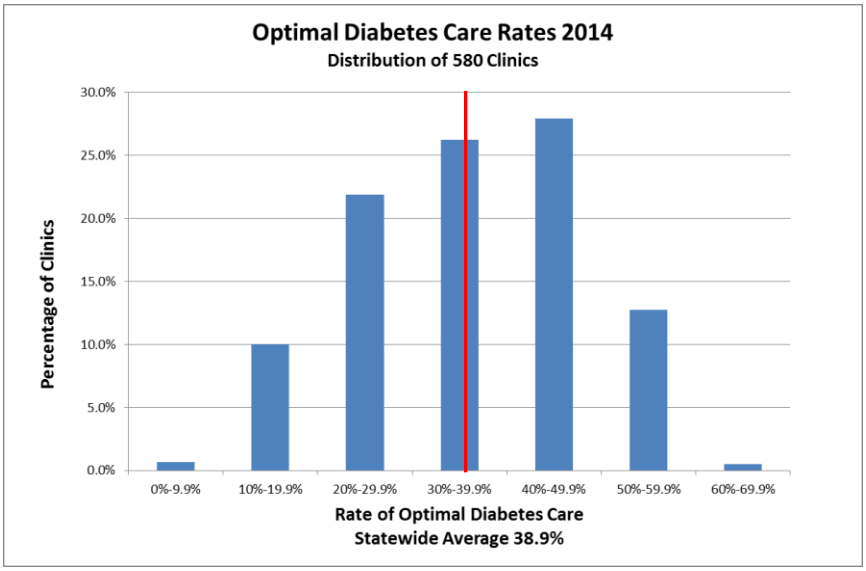
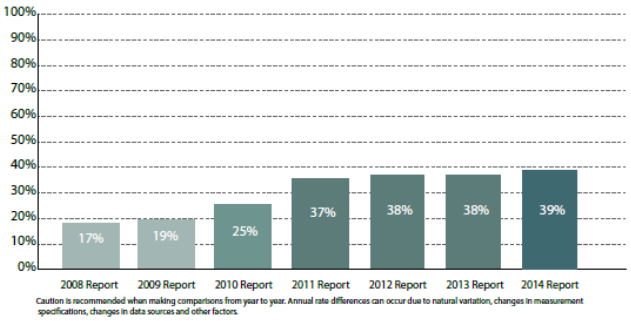


TABLE 10: STATEWIDE RATE FOR OPTIMAL DIABETES CARE

	Statewide Average (Weighted)	95% CI	Numerator (Patients who met treatment goals)	Denominator (Patients sampled)	Total Eligible
Optimal Diabetes Care	38.9%	38.7%-39.1%	90,499	230,818	237,354

FIGURE 7: STATEWIDE RATES FOR OPTIMAL DIABETES CARE OVER TIME



APPENDIX

Highest Performers in 2014 by Medical Group Type for Clinical Measures

The following tables 22-26 list the high-performing medical groups by type based on the clinical measures on which they were reported. These groups had above-average rates on most clinical measures.

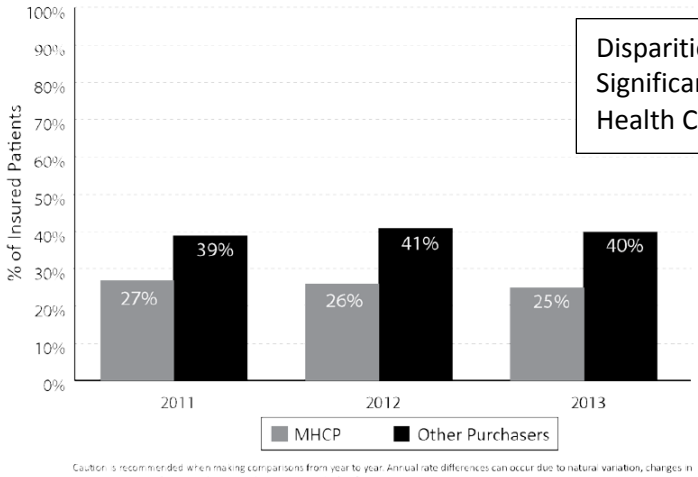
Three primary care medical groups achieved rates that were above average more frequently than others. Each of these medical groups achieved above-average rates on at least half of the clinical measures on which they were reported.

TABLE 22: HIGH PERFORMING MEDICAL GROUPS IN 2014 - PRIMARY CARE

Measure	Health Partners Clinics 13 out of 18	Park Nicollet Health Services 13 out of 18	Fairview Health Services 12 out of 18
ADHD			●
Adolescent Immunizations	●	●	
Breast Cancer Screening	●	●	●
Bronchitis	●		
Childhood Immunization Status (Combo 3)	●		●
Chlamydia Screening	●		●
Colorectal Cancer Screening	●	●	●
Controlling High Blood Pressure		●	●
COPD	●	●	●
Depression Remission at Six Months		●	
Depression Remission at Twelve Months			
Maternity Care: Primary C-Section Rate		●	
Pharyngitis	●	●	
Optimal Asthma Care-Children	●	●	●
Optimal Asthma Care-Adults	●	●	●
Optimal Diabetes Care	●	●	●
Optimal Vascular Care	●	●	●
URI	●	●	●

● = Medical group rate and CI fully above average
Blank = Measure reported but rate was average or below average.

Figure 10: Statewide Rates for Optimal Diabetes Care over Time



Caution is recommended when making comparisons from year to year. Annual rate differences can occur due to natural variation, changes in measurement specifications, changes in data sources and other factors.

Disparities Report:
Significant gaps still exist for patients under MN Health Care Programs versus all other payers

Annual Health Care Quality Report and Health Care Disparities reports available at
http://mncm.org/reports-and-websites/reports-and-data/

Methodology:

Identifying High Performing Medical Groups/Clinics

For each measure, both individual medical group rates and a medical group average rate were calculated. Medical groups that achieved high performance were identified by comparing the individual medical group/clinic rate with the medical group average. Medical groups that had rates that were fully above the medical group average and 95 percent confidence intervals were noted as high performers. Additionally, the Top 15 performers are identified.

Identifying Medical Groups and Clinics with Biggest Improvements

For each measure, individual medical group and clinic rates during report year 2014 were compared with their rates during report year 2013, 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 (2012, 2013 and 2014).
The percent and number of medical groups were reported for each of the following patterns of rate changes over the past three years for each measure:

- 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.

7.2.4.1. Confidence intervals

Confidence intervals using the method of Agresti and Coull

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

Formulas for the confidence intervals

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

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

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

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

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 outcomes and processes to reduce risk of long term complications for patients with diabetes and identifies meaningful differences among providers.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.*
If only one set of specifications for each component, 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
Note: *Applies to the overall composite measure.*
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*)

For this patient level all-or-none composite measure, elements missing from any component (e.g. visit but no blood pressure during the measurement year) are counted as a numerator component fail and therefore the patient would be accounted for and 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)

The impact of missing data on measure calculations is minimal. For 2013 dates of service on over 230,800 diabetic patients submitted for rate calculation two variables were considered 1) with in the appropriate measurement timeframe and 2) valid values submitted:

Variable	Within measure period	Invalid values
A1c	96.8%	0.003%
Blood Pressure	99.8%	0.02%
LDL	89.8%	0.9%

Tobacco Status documented – 99.8%
Aspirin or anti-platelets if IVD- 97.2% had documented aspirin or anti-platelet in the measurement year or the date of a valid contraindication.

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*)

Patients with missing data are not excluded from the measure. Elements missing from any component are counted as a numerator component fail and remain in the denominator.

_____

2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH
Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*
This composite measure is a patient level all-or-none composite in which the desired goal is for the patient is to achieve multiple intermediate physiological clinical outcome and medication use targets to best reduce their overall risk of developing long term complications (acute MI, cardiovascular and peripheral vascular disease, kidney damage and failure, loss of vision, amputation, etc.) Reducing modifiable risks was the reason why this measure was developed. The components of this measure include blood sugar and blood pressure control, being tobacco-free, appropriate use of statins and daily aspirin or anti-platelet use if ischemic vascular disease.

Achieving the intermediate physiological outcome targets related to blood pressure and glycemic control in addition being tobacco free and use of daily aspirin and statins where appropriate are the diabetic patient's best mechanisms of avoiding or postponing long term complications associated with this chronic condition which affects millions of Americans. Measuring providers separately on individual targets is not as patient centric as a measure that seeks to reduce multiple risk factors for each patient.

Diabetic patients are more likely to reduce their overall risk and maximize health outcomes by achieving several intermediate physiological targets.

A multifactorial approach to diabetes care that includes emphasis on blood pressure, lipids, glucose, aspirin use and non-use of tobacco will maximize health outcomes far more than a strategy that is limited to just one or two of these clinical domains (American Diabetes Association, 2014; Duckworth, 2009; Gaede, 2008; Holman, 2008a).

The components of this patient level all-or-none composite measure, though they can be analyzed as individual components especially for purposes of understanding opportunities within the composite measure, are treated as a whole. There is no weighting of the components; it is an all-or-none measure.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)
The components of this measure were selected as intermediate outcomes and medication use that can significantly reduce the diabetic patient's risk of developing or delaying microvascular and macrovascular or other long term complications associated with this chronic condition.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

A multifactorial approach to diabetes care that includes emphasis on blood pressure, lipids, glucose, aspirin use and non-use of tobacco will maximize health outcomes far more than a strategy that is limited to just one or two of these clinical domains (American Diabetes Association, 2014; Duckworth, 2009; Gaede, 2008; Holman, 2008a).

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible
NA

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*
NA

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)
NA

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)
NA