**National Quality Forum—Measure Testing (subcriteria 2a2, 2b1-2b6)**

**Measure Number** (*if previously endorsed*)**:** 0055

**Measure Title**: Comprehensive Diabetes Care: Eye exam (retinal) performed

**Date of Submission**: 3/5/2018

**Type of Measure:**

|  |  |
| --- | --- |
| Outcome (*including PRO-PM*) | Composite – ***STOP – use composite testing form*** |
| Intermediate Clinical Outcome | Cost/resource |
| Process *(including Appropriate Use)* | 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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** 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 (2b1-2b6) 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 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](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **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 **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **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 **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**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)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and 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. The degree of consensus and any areas of disagreement must be provided/discussed.  **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.** 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 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.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: | 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*).

N/A

**1.3. What are the dates of the data used in testing**? 2010-2012

**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.20*)** | **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*)

Health Plan Level

We calculated the measure score reliability and construct validity from HEDIS data that included 416 commercial health plans, 500 Medicare health plans, and 197 Medicaid health plans. The sample included all commercial, Medicare, and Medicaid health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size

Physician Level:

We also calculated measure score reliability from physician/practice level data from the NCQA Diabetes Recognition Program (DRP) that included 3676 physicians. Construct validity was calculated with data from a sample of 653 physicians/practices.

**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*)

2012 data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included HEDIS data collection and the median eligible population for the measure across health plans.

HEDIS Health Plan

|  |  |  |
| --- | --- | --- |
| **Product type** | **Number of plans** | **Median number of eligible patients per plan** |
| Commercial HMO | 218 | 7,433 |
| Commercial PPO | 198 | 14,513 |
| Medicaid HMO | 194 | 3,114 |
| Medicare HMO | 349 | 4,134 |
| Medicare PPO | 151 | 4,110 |

NCQA’s Diabetes Recognition Program currently has more than 10,000 clinicians in solo and group practice who hold recognition for providing quality care for their patients with diabetes. Individual clinicians or clinicians within a group practice must have face to face contact with and submit data on care delivered for a 12-month period to at least 25 different eligible adults patients with diabetes. Below is a description of the sample. It includes the number of physicians and practices reporting on this measure in the DRP program in 2012.

Physician Level

|  |  |  |
| --- | --- | --- |
| **Analysis** | **Number of physicians** | **Median denominator size** |
| Reliability | 3,676 | 25 |
| Construct Validity | 653 | 25 |

**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**.

Reliability:

Reliability of the health plan measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

Reliability of the physician/practice level measure in the DRP was tested using a beta-binomial calculation. This analysis included the entire DRP sample (described above).

Validity:

Validity of the health plan measure was demonstrated through construct validity using the entire HEDIS data sample (described above) and through a systematic assessment of face validity with expert panels.

Validity was demonstrated through construct validity using data from a sample of 653 physicians/practices and through a systematic assessment of face validity with expert panels.

**1.8** **What were the social risk factors that were available and analyzed**? For example, patient-reported data (e.g., income, education, language), proxy variables when social risk data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate) which do not have to be a proxy for patient-level data.

We did not analyze performance by social risk factors.

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

**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 testing of performance measure score:

Reliability was estimated by using the beta-binomial model for the health plan measure and physician/practice level DRP measure. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

**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*)

Health Plan Level

|  |  |
| --- | --- |
| **Product Type** | **Reliability per Beta Binomial Model** |
| Commercial | 1.0 |
| Medicare | 0.99 |
| Medicaid | 0.96 |

Physician Level

|  |  |
| --- | --- |
| **Product Type** | **Reliability per Beta Binomial Model** |
| Diabetes Recognition Program | 0.80 |

**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?*)

Health Plan Level

The values for the beta-binomial statistic across all product lines for the health plan level measure suggest the measure has high reliability.

Physician Level

The value for the beta-binomial statistic for the physician level measure suggests the measure has high reliability.

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

**2b1. VALIDITY TESTING**

**2b1.1. 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*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. 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)*

Method of Testing Construct Validity – Health Plan Level

We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality hypothesized to be related, which are listed below.

* HbA1c Testing
* Hemoglobin (HbA1c) Poor Control (>9.0%)
* Medical attention for nephropathy
* Hemoglobin (HbA1c) Control (<8%)

To test these correlations, we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 to +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable. Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Method of Testing Construct Validity – Physician Level

We tested for construct validity by exploring whether the measure was correlated with other similar measures of quality in NCQA’s Diabetes Recognition Program hypothesized to be related, which are listed below.

* Blood pressure control
* Foot exam
* Medical attention for nephropathy

We tested the correlations using the Pearson correlation test described above.

Method of Assessing Face Validity – Health Plan Level

We describe below NCQA’s process for both measure development and maintenance, which includes substantial feedback from 10 standing expert panels and 16 standing Measurement Advisory Panels, review and voting by our Committee on Performance Measurement and NCQA’s Board of Directors. In addition, all new measures and measures undergoing significant revision are included in our annual HEDIS 30-day public comment period, which on average receives over 800 distinct comments from the field including organizations that are measured by NCQA, providers, patients, policy makers and advocates. NCQA refines our measures continuously through feedback received from our Policy Clarification (PCS) Web Portal, which on average receives and responds to over 3,000 inquiries each year. All HEDIS measures are audited by certified firms according to standards, policies and procedures outlined in HEDIS Volume 7. Combined, these processes which NCQA has used for over 25 years assure that the measures we use are valid.

NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle for all plan-level HEDIS measures.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) 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 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 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.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publicly reported and may be used for scoring in accreditation.

STEP 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reviewed periodically, based on changes in evidence and guidelines. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA’s Policy Clarification Support (PCS) 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. Over the past four years, NCQA has received and responded to an average of 39 inquiries per year on this measure.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year’s data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year’s HEDIS Volume 2 and in other relevant NCQA programs.

Method of Assessing Face Validity - Physician Level

The physician level measure was tested for face validity with four panels of experts. The Diabetes Recognition Program (DRP) Advisory Committee included 7 experts in diabetes care including representation by clinicians, health plans, integrated health systems and research organizations; DMAP, CPM and the Clinical Programs Committee (CPC). NCQA’s CPC’s oversees the evolution of NCQA’s recognition programs and related measures including the Diabetes Recognition Program, the Heart/Stroke Recognition Program, the Patient Centered Medical Home and Patient-Centered Specialty Practice Recognition Program, among others. The CPC includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 18 members. The CPC is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of clinical recognition programs. CPC members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

Construct Validity – Health Plan Level

The results from construct validity testing of the health plan level measure are presented by product line in Tables 1a, 1b, and 1c below.

Table 1a. Correlations among Diabetes Measures in Commercial Health Plans - 2012

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Pearson Correlation Coefficient** | | | |
| HbA1c Testing | HbA1c Poor Control (>9.0%) | Medical Attention for Diabetic Nephropathy | HbA1c Good Control (<8.0%) |
| CDC - Eye Exams | 0.69 | -0.63 | 0.72 | 0.65 |

Note: All correlations are significant at p<0.0001

Table 1b. Correlations among Diabetes Measures in Medicaid Health Plans - 2012

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Pearson Correlation Coefficient** | | | |
| HbA1c Testing | HbA1c Poor Control (>9.0%) | Medical Attention for Diabetic Nephropathy | HbA1c Good Control (<8.0%) |
| CDC - Eye Exams | 0.53 | -0.53 | 0.45 | 0.51 |

Note: All correlations are significant at p<0.0001

Table 1c. Correlations among Diabetes Measures in Medicare Health Plans - 2012

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Pearson Correlation Coefficient** | | | |
| HbA1c Testing | HbA1c Poor Control (>9.0%) | Medical Attention for Diabetic Nephropathy | HbA1c Good Control (<8.0%) |
| CDC - Eye Exams | 0.60 | -0.52 | 0.38 | 0.51 |

Note: All correlations are significant at p<0.0001

Construct Validity – Physician Level

Table 2a below provides the results from construct validity testing of the physician level measure.

Table 2a. Correlations among Diabetes Measures in the NCQA Diabetes Recognition Program - 2012

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Pearson Correlation Coefficient** | | |
| Blood Pressure Control | Medical Attention to Nephropathy | Foot Exam |
| CDC – Eye Exams | 0.41 | 0.26 | 0.42 |

Note: All correlations are significant at p<0.0001

**2b1.4. 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?*)

Construct Validity – Health Plan Level

Across all product lines, the correlations are moderate to strong and statistically significant. These results confirmed the hypothesis that the diabetes measures are correlated with each other. Coefficients with absolute value of less than .3 are generally considered indicative of weak associations. Absolute values of .3 to .59 are considered moderate associations, absolute values of .6 to .69 indicate a strong positive relationship, and absolute values of .7 or higher indicate a very strong positive relationship. These correlation results suggest that at the plan level the measure has sufficient validity.

*Note: Correlation values with the HbA1c Poor Control measure are all negative because it is a “lower is better quality” measure, while the other measures are all "higher is better".*

Construct Validity - Physician Level

At the physician level, the *CDC-Eye Exam* measure has a moderate correlation with the *Blood Pressure Control* and *Foot Exam* measures in the Diabetes Recognition Program. The correlation between the *Eye Exam* measure and the *Medical Attention to Nephropathy* measure is lower and indicates a slightly weaker association. Overall these correlation results suggest that the physician level measure has sufficient validity.

Face Validity – Health Plan Level

NCQA’s expert panels, our measurement advisory panels and our Committee on Performance Measurement agreed that the *CDC – Eye Exams* measure is measuring what it intends to measure. The results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across health plans.

Face Validity – Physician Level

The results indicate that the multiple experts, stakeholders and NCQA’s Clinical Programs Committee concluded with good agreement that the measure as specified is measuring what it intends to measure and that the results of the measurement allow users to make the correct conclusions about the quality of care that is provided and will accurately differentiate quality across providers.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

**2b2.1. 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*)

Testing was not performed for the excluded sample.

**2b2.2. 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*)

Testing was not performed for the excluded sample.

**2b2.3. 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*)

Testing was not performed for the excluded sample.

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

**2b3. 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*** [***2b4***](#section2b5)***.***

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

**2b3.1.1 If using a statistical risk model, provide detailed risk model specifications, including the risk model method, risk factors, coefficients, equations, codes with descriptors, and definitions.**

N/A

**2b3.2. 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**.   
N/A

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk 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*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors?  
N/A

**2b3.3b. How was the conceptual model of how social risk impacts this outcome developed? Please check all that apply:**

**Published literature**

**Internal data analysis**

**Other (please describe)**

**2b3.4a. What were the statistical results of the analyses used to select risk factors?**N/A

**2b3.4b. Describe the analyses and interpretation resulting in the decision to select social risk factors** *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects.)* **Also describe the impact of adjusting for social risk (or not) on providers at high or low extremes of risk.**

N/A

**2b3.5. 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*)  
N/A

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b3.9. Results of Risk Stratification Analysis**:

**2b3.10. 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*)

**2b3.11.** **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*)

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

**2b4. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b4.1. 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)*

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each measure. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample, size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans performance is significantly different from each other.

**2b4.2. 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*)

Health Plan Level - 2012

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Product Type** | **N** | **Mean (%)** | **St Dev (%)** | **P10th (%)** | **P25th (%)** | **P50th (%)** | **P75th (%)** | **P90th (%)** | **IQR (%)** | **P value** |
| Commercial HMO | 218 | 56.82 | 13.83 | 38.44 | 47.69 | 57.58 | 66.83 | 75.13 | 19.14 | <0.05 |
| Commercial PPO | 198 | 48.80 | 10.03 | 33.20 | 43.08 | 49.85 | 54.74 | 60.25 | 11.66 | <0.05 |
| Medicaid HMO | 194 | 53.22 | 11.96 | 37.14 | 44.37 | 54.43 | 62.46 | 67.64 | 18.09 | <0.05 |
| Medicare HMO | 349 | 66.79 | 11.40 | 53.04 | 59.85 | 67.35 | 74.35 | 80.87 | 14.60 | <0.05 |
| Medicare PPO | 151 | 64.63 | 10.34 | 53.04 | 56.94 | 64.72 | 70.56 | 77.86 | 13.62 | <0.05 |

N = total number of plans reporting data

IQR: Interquartile range

p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

Physician Level - 2012

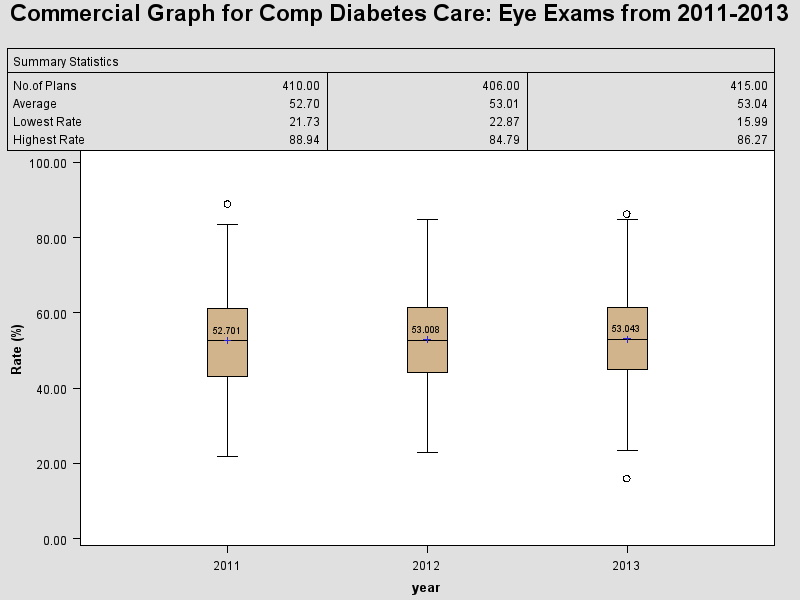
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **N (# of clinicians)** | **Mean (%)** | **St Dev**  **(%)** | **10th**  **(%)** | **25th**  **(%)** | **50th**  **(%)** | **75th**  **(%)** | **90th**  **(%)** | **IQR**  **(%)** | **p value** |
| 3676 | 52.0 | 25.6 | 16.0 | 32.0 | 53.0 | 72.0 | 87.0 | 40.0 | <0.05 |

IQR: Interquartile range

p-value: p value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile

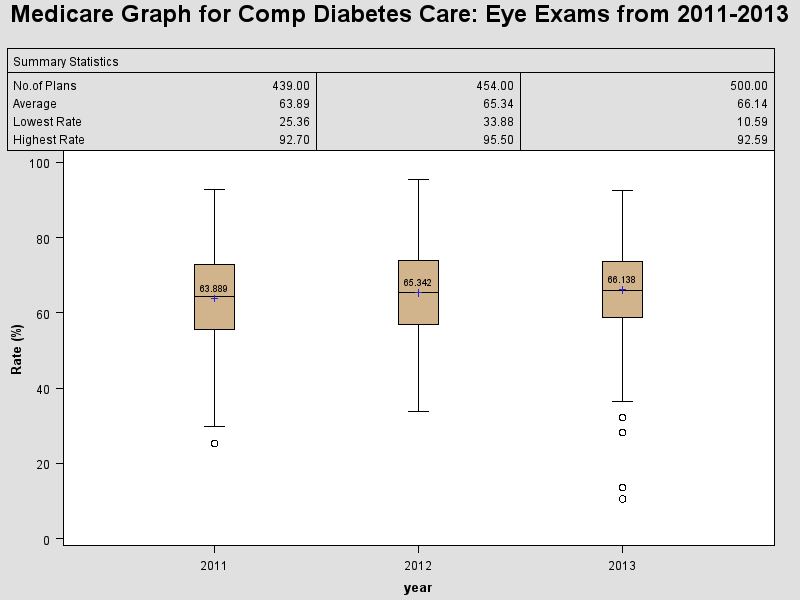
Health Plan

**Chart 1. Boxplot of Eye Exams Measure, Commercial, HEDIS 2011-2013\***



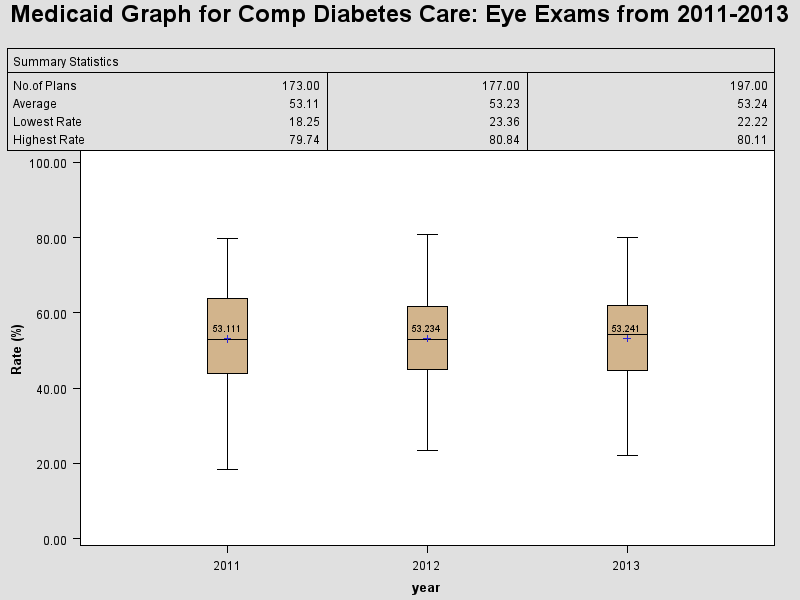
\* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

**Chart 2. Boxplot of Eye Exams Measure, Medicare, HEDIS 2011-2013\***



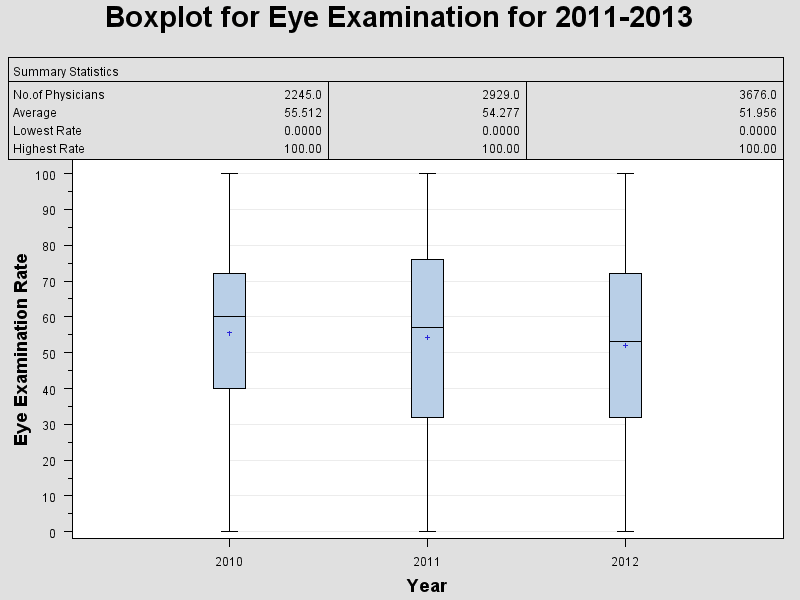
\* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

**Chart 3. Boxplot of Eye Exams Measure, Medicaid HEDIS 2011-2013\***



\* In this chart data is presented in HEDIS reporting years, which are a year ahead of the measurement year. Therefore, the measurement year is 2010-2012

Physician Level

**Chart 4. Boxplot of Eye Exams Measure, Diabetes Recognition Program, 2010-2012**

**2b4.3. 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?*)

Health Plan Level

Across all product lines, the difference between the 25th (better performance) and 75th percentile is statistically significant. Overall, these results suggest there are meaningful differences in performance.

Physician Level

The difference between the 25th and 75th percentile is statistically significant, suggesting there are meaningful differences in performance.

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

**2b5. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This item is directed to measures that are risk-adjusted (with or without social risk factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or 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).* ***Comparability is not required when comparing performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.***

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

N/A

**2b5.2. 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*)

N/A

**2b5.3. What is your interpretation of the results in terms of the differences in 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*)  
N/A

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

**2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b6.1. 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*)

This measure is collected with a complete sample.

**2b6.2. 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*)

This measure is collected with a complete sample.

**2b6.3. 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*)

This measure is collected with a complete sample.