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

**Measure Number** (*if previously endorsed*)**:** 0059

**Measure Title**: Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%)

**Date of Submission**: 8/1/2019

**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: 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*).

N/A

**1.3. What are the dates of the data used in testing**? Click here to enter date range

Testing of performance measure score with beta binomial reliability and testing of construct validity with the Pearson Correlation were performed using HEDIS 2019 plan level data, measurement year 2018.

**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*)

This measure assesses whether adults enrolled in commercial, Medicare, and Medicaid plans who have diabetes (type 1 and type 2) had their most recent HbA1c level greater than 9.0%. Therefore, testing was done at the health-plan level, which is appropriate for the level of reporting for this measure.

Note: A lower rate indicates better performance for this measure (i.e., low rates of poor control indicate better care).

We calculated the measure score reliability and construct validity from HEDIS data that included 378 commercial health plans, 477 Medicare health plans, and 241 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.

**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*)

Table 1 below provides a description of the data submitted for 2018, including the median denominator size per plan. Data are summarized at the health plan level and stratified by plan type (i.e. commercial, Medicaid, Medicare). Since data can be collected and reported from two data sources (administrative claims and medical record review), the vast majority of plans use a combination of data from administrative claims data and a sample of 411 of medical records they review to report their performance rates.

Table 1. Commercial, Medicaid, and Medicare plans reporting the Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%), 2018.

|  |  |  |
| --- | --- | --- |
| Product Type | Number of Plans | Median Denominator Size/Plan |
| Commercial | 378 | 411 |
| Medicaid | 241 | 411 |
| Medicare | 477 | 411 |

**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).

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.

**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 social risk factors. This measure of health plan performance is specified to be reported separately by commercial, Medicaid and Medicare plan types, which serves as a proxy for income and other socioeconomic 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 was estimated by using the Beta-binomial model (Adams, 2009) for this health plan measure. Beta-binomial is appropriate for estimating the reliability of pass/fail rate measures. 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.

Adams, J.L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

**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*)

Table 2 provides the reliability for the overall measure as shown by the Beta-binomial model as well as the distribution of individual plan reliability.

Table 2. Overall Beta-binomial statistic and distribution of plan reliability for commercial, Medicaid, and Medicare product lines, 2018

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Product Line | Overall Reliability | Min | Percentiles | | | | | Max |
| 10th | 25th | 50th | 75th | 90th |
| Commercial | 0.996 | 0.831 | 0.980 | 0.982 | 0.983 | 0.987 | 0.992 | 1.000 |
| Medicaid | 0.983 | 0.627 | 0.915 | 0.955 | 0.966 | 0.973 | 0.977 | 1.000 |
| Medicare | 0.980 | 0.792 | 0.970 | 0.975 | 0.977 | 0.982 | 0.985 | 1.000 |

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

The values for the beta-binomial statistic across all product lines for the health plan level measure are all greater than 0.7, indicating the measure has very good reliability. The 10-90th percentile distribution of health plan level-reliability on this measure show the vast majority of health plans exceeded the minimally accepted threshold of 0.7, and the majority of plans exceeded 0.9. Strong reliability is demonstrated since the majority of variance is due to signal and not to noise.

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

**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)*

We tested for construct validity of the Comprehensive Diabetes Care (CDC): HbA1c Poor Control (>9.0%) measure by exploring whether it was correlated with other similar measures of quality hypothesized which are listed below.

* CDC: Hemoglobin A1c (HbA1c) Testing: The percentage of adults 18-75 with diabetes that had an HbA1c test performed during the measurement year.
* CDC: HbA1c Control (<8.0%): The percentage of adults 18-75 with diabetes whose most recent HbA1c level is <8% during the measurement year.
* CDC: Eye Exam (Retinal) Performed: The percentage of adults 18-75 with diabetes that had an eye screening for diabetic retinal disease during the measurement year.
* CDC: Medical Attention for Nephropathy: The percentage of adults 18-75 with diabetes that had a nephropathy screening test or evidence of nephropathy during the measurement year.
* CDC: Blood Pressure Control (<140/90 mm Hg): The percentage of adults 18-75 with diabetes whose most recent blood pressure level taken during the measurement year is <140/90 mm Hg.

These measures were chosen for construct validity because they are similarly focused on a population with diabetes (type 1 and type 2) and focus on evidenced-based monitoring and treatment for patients with diabetes. We hypothesized that a plan that does well on these measures for diabetes would also do well on this blood pressure control measure for patients who have diabetes.

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.

*\* Note: All HEDIS value sets are updated annually with the most current codes available. The information below details the process we used to convert value sets that used ICD-9 codes to ICD-10 codes in 2015. \**

ICD-10 CONVERSION:

In preparation for the national implementation of ICD-10 in 2015, NCQA conducted a systematic mapping of all value sets maintained by the organization to ensure the new values used for reporting maintained the reliability, validity, and intent of the original specification.

Steps in ICD-9 to ICD-10 Conversion Process

1. NCQA first identified value sets within the measure that included ICD-9 codes. We used General Equivalence Mapping (GEM) to identify ICD-10 codes that map to ICD-9 codes and reviewed GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA then searched for additional codes (not identified by GEM mapping step) that should be considered due to the expansion of concepts in ICD-10. Using ICD-10 tabular list and ICD-10 Index, searches by diagnosis or procedure name were conducted to identify appropriate codes.
3. NCQA HEDIS Expert Coding Panel review: Updated value set recommendations were presented for expert review and feedback.
4. NCQA RMAP clinical review: Due to increase specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is consistent and appropriate given the scope of the measure.
5. New value sets containing ICD-10 code recommendations were posted for public review and comment in 2014 and updated in 2015. Comments received were reconciled with additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
6. NCQA staff finalized value sets containing ICD-10 codes for publication in 2015.

Tools Used to Identify/Map to ICD-10

All tools used for mapping/code identification from CMS ICD-10 website

(http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html).

GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

*Expert Participation*

The NCQA HEDIS Expert Coding Panel reviewed and provided feedback on staff recommendations. Names and credentials of the experts who served on these panels are listed under Additional Information, Ad. 1. Workgroup/Expert Panel Involved in Measure Development.

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

The results from construct validity testing of the health plan level measure are presented by product line in Tables 3a, 3b, and 3c below.

Note: A lower rate indicates better performance for this measure (i.e., low rates of poor control indicate better care), which explains the negative relationship with the other measures, for which a higher rate indicates better performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3a. Correlations among Diabetes Measures in Commercial Health Plans, 2018. | | | | | |
|  | **Pearson Correlation Coefficients** | | | | |
| HbA1c Testing | HbA1c Control (<8.0%) | Eye Exams | Medical Attention for Diabetic Nephropathy | Blood Pressure Control <140/90 |
| HbA1c Poor Control >9.0%) | -0.3266 | -0.9896 | -0.4370 | -0.3419 | -0.8986 |

Note: All correlations are significant at p<0.0001

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3b. Correlations among Diabetes Measures in Medicaid Health Plans, 2018. | | | | | |
|  | **Pearson Correlation Coefficients** | | | | |
| HbA1c Testing | HbA1c Control (<8.0%) | Eye Exams | Medical Attention for Diabetic Nephropathy | Blood Pressure Control <140/90 |
| HbA1c Poor Control >9.0%) | -0.6663 | -0.9868 | -0.6275 | -0.3286 | -0.7820 |

Note: All correlations are significant at p<0.0001

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3c. Correlations among Diabetes Measures in Medicare Health Plans, 2018. | | | | | |
|  | **Pearson Correlation Coefficients** | | | | |
| HbA1c Testing | HbA1c Control (<8.0%) | Eye Exams | Medical Attention for Diabetic Nephropathy | Blood Pressure Control <140/90 |
| HbA1c Poor Control (>9.0%) | -0.5984 | -0.9657 | -0.6049 | -0.4761 | -0.5914 |

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?*)  
Across all product lines, the correlations are moderate to very strong and statistically significant. These results confirm the hypothesis that plan performance on these 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 this plan level measure has sufficient validity.

*Note: The correlation values are all negative because the HbA1c Poor Control measure is a “lower is better quality” measure, while the other measures are “higher is better”. This indicates that plans that have low rates on this measure will have high rates on the others.*

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

**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*)

We did not perform testing of the following exclusions for this submission:

* Gestational diabetes
* Steroid-induced diabetes

NCQA engaged expert panels to inform the face validity of these exclusions for this measure, which aligns with evidence focused on the general population of people with Type I or Type II diabetes. This measure has been reviewed by NCQA’s Diabetes Measurement Advisory Panel, Cardiovascular Measurement Advisory Panel, Technical Measurement Advisory Panel, and the Committee on Performance Measurement. The measure also received public comment feedback upon initial development.

Hospice, I-SNPs and Long-Term Care Institutions

These exclusions were also not formally tested for this submission. This measure is designed to be scientifically valid and feasible for comparing the quality of care provided to general populations, such as healthy older adults or those with a single condition. Patients receiving hospice, enrolled in an I-SNP, or residing in a long-term care institution would likely have different care needs and quality concerns, therefore they are excluded from this measure.

Advanced Illness and Frailty

For HEDIS 2019 (measurement year 2018), NCQA added exclusions for advanced illness and frailty to the Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%) measure. NCQA decided to explore implementing these exclusions, recognizing that for individuals with limited life expectancy, advanced illness and frailty, the focus of this measure may not be clinically appropriate, relevant or in line with the patient’s goals of care. We performed a review of literature on different approaches to defining advanced illness and used this, along with feedback received from expert work groups, measurement advisory panels and public comment to create a list of illnesses, conditions and service codes to be included in testing. The conditions included: dementia and other neurodegenerative conditions, emphysema, end stage renal disease (ESRD), heart failure, liver failure, metastatic cancer, pulmonary fibrosis and respiratory failure.

NCQA then conducted a search of ICD-10 codes that were relevant to each of the conditions to create value sets for testing. To identify those with dementia, NCQA also included drug codes for medications such as donepezil hydrochloride and galantamine hydrobromide, to capture those who may not carry a diagnosis of dementia but are prescribed a drug for treatment.

The proxy for frailty was developed based on previously studied approaches[[1]](#footnote-2), [[2]](#footnote-3), [[3]](#footnote-4) and feedback received from expert work groups and measurement advisory panels. The proxy is comprised of HCPCS and ICD-10 codes for diagnoses or services that can indicate when an individual is frail or dependent in activities of daily living. Examples include: gait abnormality, abnormal loss of weight and underweight, adult failure to thrive, debility, fall, pressure ulcer, durable medical equipment (hospital bed, walker, portable or home oxygen, wheelchair), bed confinement, palliative care and age-related physical debility. Members met the frailty proxy criteria if they had a claim for any of the codes included in the frailty code set in the measurement year.

To determine the feasibility and impact of applying this exclusion to the measure, NCQA used a research database that consisted of two years of inpatient, outpatient, and pharmacy claims for members age 18 and older enrolled in a sample of Medicare Advantage plans (N=25). NCQA compared several approaches for identifying the advanced illness and frailty populations, examining different age ranges and diagnosis positions and their impact on the denominator. The results of those queries along with input from the expert work groups, measurement advisory panels and public comment led us to determine that the best approach for identifying the advanced illness and frailty population that should be excluded from the measure was to apply the following criteria:

* Adults 66 and older as of December 31 of the measurement year (all product lines) with frailty and advanced illness

**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*)

Table 4 shows the results of applying the exclusion of adults 66 and older with advanced illness and frailty to the Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%) measure.

Table 4. Impact of applying the advanced illness and frailty for patients aged 66 and older

|  |  |  |
| --- | --- | --- |
| Number of Plans  (N) | Average Number Excluded | Average % Removed by Exclusion |
| 25 | 350 | 2.0 |

**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*)

Advanced Illness and Frailty

The advanced illness and frailty exclusion had a small impact on the eligible population: 2.0% on average were removed for advance illness and frailty. Feedback from NCQA’s expert work groups and measurement advisory panels, as well as public comment feedback, supported the application of this exclusion to the Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%) measure for clinical reasons. By implementing this exclusion, those providing care to patients with advanced illness and frailty can focus on care that is more appropriate for their conditions and health status. Attention can be more focused on quality measures that capture services and care processes that are most relevant for this population (e.g., improving care transitions, getting follow-up after acute care episodes, or avoiding preventable hospitalizations).

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

**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**.

NCQA recognizes that there is a growing body of literature that might support risk adjustment or stratification of intermediate outcome measures. However, at this time, NCQA does not currently risk adjust this measure given the potential to mask poor performance and disparities in care.

NCQA conducted a study on the Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%) measure among Medicare Advantage plans to assess whether to account for a member’s socioeconomic status (SES) when comparing plan performance. A qualitative assessment included key informant interviews exploring ways in which SES may affect performance on this and other select HEDIS measures, and whether there was a conceptual basis for case-mix adjustment or other strategies. In the quantitative analysis, we assessed whether SES affected plan performance, using member low-income status, dual eligibility, and disability as proxies for SES. For this measure, adjusting for SES did not have a meaningful impact on results. When adjusting for disparity in performance between low- and high-SES populations, plan ranks were not substantially impacted. When accounting for clinical and demographic factors, we found that low-SES beneficiaries were as likely, or more likely, to receive recommended care as high-SES beneficiaries. Our results suggest there is neither a conceptual nor empirical basis for risk adjustment for this measure.

**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 performance of the two plans differs significantly.

**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*)

*Note that for this measure, a lower rate equals better performance so the values in the 10th percentile are representative of the top performers and the values in the 90th percentile are representative of the poor performers.*

Table 5. Variation in Performance for commercial, Medicaid, and Medicare health plans, 2018.

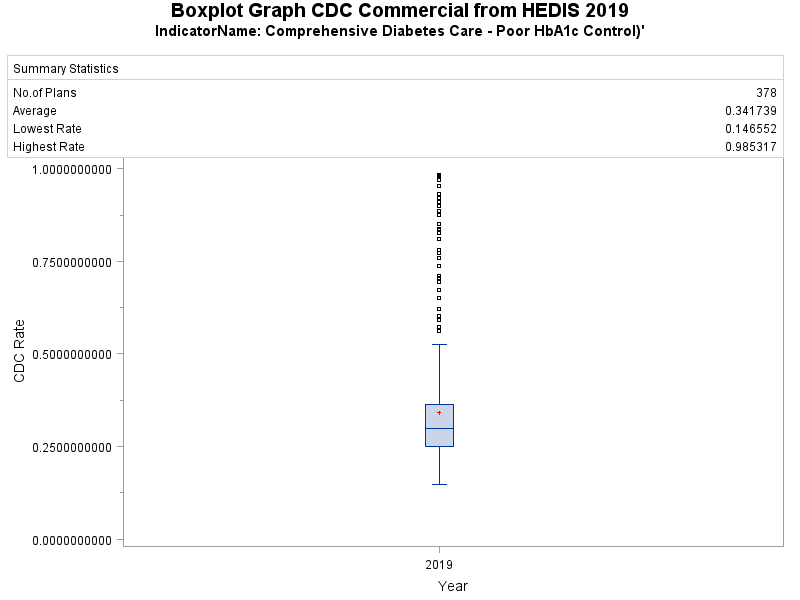
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Plan Type** | **N** | **Average (%)** | **St Dev**  **(%)** | **10th**  **(%)** | **25th**  **(%)** | **50th**  **(%)** | **75th**  **(%)** | **90th**  **(%)** | **IQR**  **(%)** | **p-value** |
| Commercial | 378 | 34.17 | 16.20 | 21.17 | 25.06 | 29.93 | 36.48 | 49.31 | 11.42 | <0.0001 |
| Medicaid | 241 | 41.15 | 13.52 | 27.98 | 32.85 | 38.44 | 46.53 | 57.91 | 13.69 | <0.0001 |
| Medicare | 477 | 21.86 | 12.24 | 11.39 | 13.87 | 18.25 | 25.35 | 36.40 | 11.48 | <0.0001 |

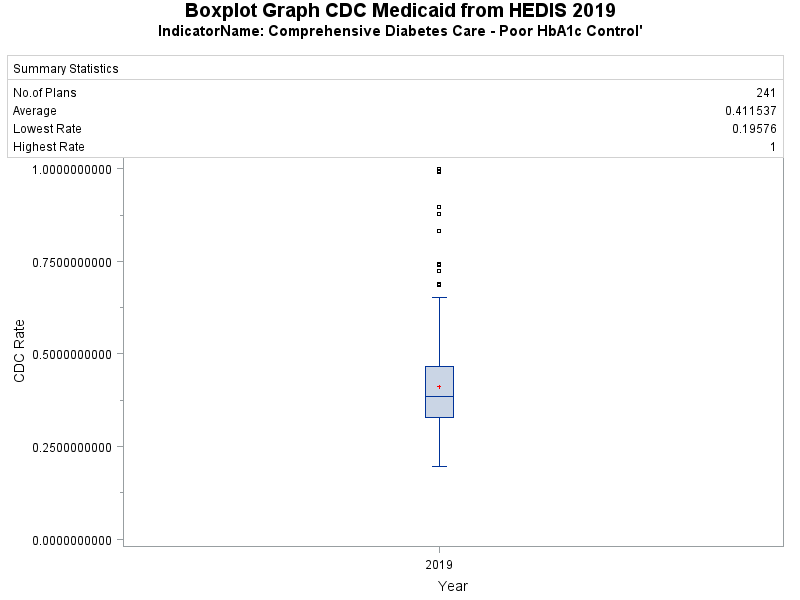
N = total number of plans reporting data

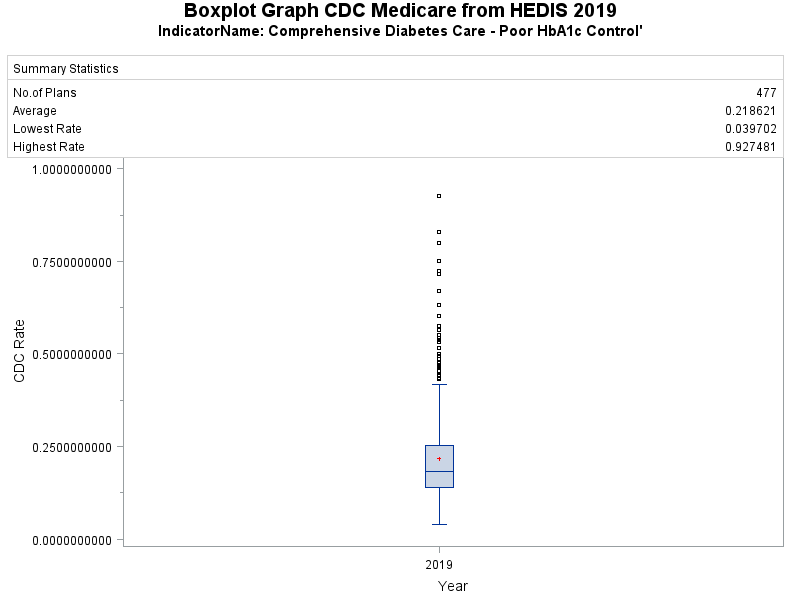
IQR: Interquartile range

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

Box plots for HEDIS 2019 (Measurement year 2018) Variation in Performance Across Health Plans are included below for your reference.



****

****

**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?*)

The results above indicate there is meaningful difference in performance. Across all product lines, the difference between the 25th (better performance) and 75th percentile is statistically significant.

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

**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.***

The Comprehensive Diabetes Care: HbA1c Poor Control (>9.0%) measure has only one set of specifications.

**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*)

HEDIS measures apply to enrolled members in a health plan, and NCQA has a rigorous audit process to ensure the eligible population and numerator events for each measure are correctly identified and reported. The audit process is designed to verify primary data sources used to populate measures and ensure specifications are correctly implemented.

The HEDIS Compliance Audit addresses the following functions:

- Information practices and control procedures

- Sampling methods and procedures

- Data integrity

- Compliance with HEDIS specifications

- Analytic file production

- Reporting and documentation

**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*)

HEDIS addresses missing data in a structured way through its audit process. HEDIS measures apply to enrolled members in a health plan, and NCQA-certified auditors use standard audit methodologies to assess whether data sources are missing data. If a data source is found to be missing data, and the issues cannot be rectified, the auditor will assign a “materially biased” designation to the measure for that reporting plan, and the rate will not be used. Once measures are added to HEDIS, NCQA conducts a first-year analysis to assess the measure’s feasibility once widely implemented in the field. This analysis includes an assessment of how many plans report valid rates vs. rates that are materially biased (or have other issues, such as small denominators). These considerations are weighed in the deliberation process before measures are approved for public reporting.

**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*)

The denominator of this measure is identified using claims data and not subject to difference between response or nonresponse. This measure goes through the NCQA audit process each year to identify potential errors or bias in results. Only performances rates that have been reviewed and determined not to be “materially biased” are reported and used.

1. Faurot, K.R., Funk, M.J., Pate, V., Brookhart, M.A., Patrick, A., Hanson, L.C., Castillo, W.C., Stürmer, T. 2015. Using Claims Data to Predict Dependency in Activities of Daily Living as a Proxy for Frailty. Pharmacoepidemiology and Drug Safety. 24(1): 59-66. [↑](#footnote-ref-2)
2. Segal, J.B., Chang, H.Y., Du, Y., Walston, J.D., Carlson, M.C., Varadhan, R. 2017. Development of a

   Claims-Based Frailty Indicator Anchored to a Well-Established Frailty Phenotype. Medical Care. 55(7): 716-722. [↑](#footnote-ref-3)
3. Davidoff A.J., A. Hurrida, I.H. Zuckerman, S.M. Lichtman, N. Pandya, A. Hussain, F. Hendrick, J.P. Weiner, X. Ke, M.J. Edelman. 2013. A Novel Approach to Improve Health Status Measurement in Observational Claims-Based Studies of Cancer Treatment and Outcomes. J Geriatr Oncol. 4(2):157–165. [↑](#footnote-ref-4)