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

**Measure Number** (*if previously endorsed*)**:** Click here to enter NQF number

**Measure Title**: Retesting of Abnormal Blood Glucose in Patients with Prediabetes

**Date of Submission**: 1/6/2020

**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 sub criteria 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*).

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

The measurement period (data collected from patients seen) was 8/1/2018 through 09/30/2019.

**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*)

Two sites were recruited and identified to collect data for testing and analysis at the data element level, reviewing individual patient records from the EHR and comparing to a manual review of the same cases. Testing was completed using a convenience sample, whereas sample size requirements were calculated based on estimated rates for each measure by site using a calculator based on the calculation defined by Donner-Eliasziw (see section 1.6 below).

Using specifications defined by the measure developer, both testing sites were able to access and test the critical data elements that included all components of the numerator, all components of the denominator, and all components of the exclusions. Testing was completed at the data element level and was completed on all patient cases in the sample. It should be noted that although the measure is specified at the physician and physician group level, testing was completed at the individual data element level (as opposed to signal to noise), so therefore there would not be counts of physicians included in the analysis.

* Test Site #1:  An ambulatory facility in South Carolina, part of a larger health system comprised of 8 inpatient hospitals and more than 100 outpatient facilities This facility uses Epic EHR in their facility.
* Test Site #2: An ambulatory facility in South Carolina, part of a larger system comprised of a 1,600+ bed comprehensive integrated health system, serving 1 million patients. This facility uses Cerner EHR in their facility.

**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*)

As described in 1.5, we recruited and subcontracted with two sites to collect patient-level EHR data for the measure. Patient cases were included in the testing and analysis that met the following criteria:

* Adults ages 18 and over
* All races
* All genders

Due to the large size of the data available at each site, a sample of the patients at each site were identified through a validated process as described below.

Sample size requirements were calculated based on estimated rates for each measure by site using a calculator based on the calculation defined by Donner-Eliasziw[[1]](#footnote-1). A discussion and application of the use of the kappa statistic in reliability studies is available in Sim and Wright, 2005.[[2]](#footnote-2) These methods were instituted in order to ensure that reliability testing and analyses occur on data sets that have a large enough sample size to detect statistically significant differences, thus minimizing variation due to the play of chance.

The important variables for the kappa sample size calculation are as follows:

* The value for the expected proportion of positive ratings for the measure being tested could be based on available data on the average performance of clinicians on the measure. If the average performance is 90%, the proportion of positive ratings is 0.90.
* The standard assumptions for testing projects are to specify the 2-tailed test at 80% power required to detect a difference between the value of the calculated kappa statistic and the null value for kappa, for example a kappa of .090 versus the null value of kappa of 0.60. This tests whether the difference in the kappa values of 0.30 (0.6 versus 0.9) is significant.

Each site provided us with preliminary counts of patients meeting the numerator and denominator to be used in sample size calculations. Following is a table that displays the data reported from the sites, the recommended sample size from the sample calculator, and the actual sample size for which the site was asked to collect data.

|  |  |  |
| --- | --- | --- |
|  |  | |
|  | **Site 1** | **Site 2** |
| **Numerator (preliminary counts)** | 76 | 2688 |
| **Denominator (preliminary counts)** | 260 | 5567 |
| **Calculated Sample Size** | 67 | 55.8 |
| **Actual Sample Used for Testing** | **100** | **75** |

**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 and validity of the data elements and exclusion testing utilized the same data from the practice site’s respective EHR systems of Epic and Cerner. Risk adjustment and stratification were not applied and not applicable for these measures.

**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.

There were no social risk factors accounted for as elements in the measures. However, the Supplemental Data Elements in the measure specifications include language, race, ethnicity, and payor as elements that can be collected for each measure to allow for the stratification of measure results by these variables to assess disparities and initiate subsequent quality improvement activities.

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

**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*)

By current NQF standards, data element validity testing results may be reported for reliability results. Testing description, results, and interpretation of results are given here and in 2b1. Validity Testing.

Data element validity testing was conducted utilizing Parallel Forms Reliability Testing methodology. Parallel forms reliability testing considers the analysis of agreement, through assessing the extent to which multiple formats or versions of data abstraction yield the same results. Verification of the data elements was obtained through automated data search strategies against a reference strategy (considered the gold standard) for obtaining the data elements. Manual review of the data elements was used as the reference strategy against which automated data search and extraction strategies were evaluated.

For this electronic clinical quality measure (eCQM), testing was used to determine if data elements found through electronic data pulls could be confirmed by manual abstraction of the same data elements. Testing at the level of the data elements allows for the analysis of each individual required data element included in the performance measure.

Interrater reliability (Cohen’s Kappa coefficient) is used to assess the reliability of the measure based on results from two independent reviewers trained in the same way reviewing the same patient record.  To perform inter-rater reliability testing we created an electronic data collection tool and trained the reviewers (raters) on its’ use. The reviewers separately reviewed every sampled patient and collected all data elements necessary for computation of the performance measure (contained on the electronic data collection tool). Data received was analyzed using SAS to calculate frequencies, level of agreement, and agreement statistics (Cohen’s Kappa). Cohen’s Kappa coefficient measures inter-rater agreement for qualitative items and takes into account any agreement occurring by chance.

The following table displays the interpretation of the kappa statistic:

**Kappa Strength of Agreement[[3]](#footnote-3)**  
0.00 Poor  
0.01 - 0.20 Slight  
0.21 - 0.40 Fair  
0.41 - 0.60 Moderate  
0.61 - 0.80 Substantial  
0.81 - 0.99 Almost Perfect

**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*)

The data element validity testing results including Kappa scores are presented below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **% Agree** | **Kappa** | **N** |
|  |  |  |  |  |
| **Site 1** | **Denominator** | 98 | 0.955 | 100 |
| **Numerator** | 100 | 1.000 | 86 |
|  |  |  |  |  |
| **Site 2** | **Denominator** | 83 | 0.504 | 72 |
| **Exclusions** | 76 | 0.511 | 67 |
| **Numerator** | 100 | \*\* | 23 |

\*\*Kappa scores not calculable with multiple non-responses by raters (i.e., all No/No or all Yes/Yes)

**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?*)  
For this measure, we find moderate to high levels of agreement for all data elements. Kappa scores ranged from .50 to 1.0, which is considered moderate to almost perfect.

We found a few instances across the measure where more full and accurate information could be found in the manual abstraction process than through electronic reporting. For example, if height is not measured and weight is measured at an office visit, then BMI may not be calculated and reported in a discrete field. In these cases, BMI can be found only by looking at the chart. Additionally, for exclusions related to the identification of comfort measures, hospice care ambulatory, and palliative care are not always stored in discrete fields, even if these are available. Some information is stored in comments/notes in the EHR.

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

**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)*

This measure was tested for data element validity testing, content validity, face validity, and feasibility of the data elements.

Data element validity testing was conducted utilizing Parallel Forms Reliability Testing methodology. Parallel forms reliability testing considers the analysis of agreement, through assessing the extent to which multiple formats or versions of data abstraction yield the same results. Verification of the data elements was obtained through automated data search strategies against a reference strategy (considered the gold standard) for obtaining the data elements. Manual review of the data elements was used as the reference strategy against which automated data search and extraction strategies were evaluated.

For this electronic clinical quality measure (eCQM), testing was used to determine if data elements found through electronic data pulls could be confirmed by manual abstraction of the same data elements. Testing at the level of the data elements allows for the analysis of each individual required data element included in the performance measure.

Interrater reliability (Cohen’s Kappa coefficient) is used to assess the reliability of the measure based on results from two independent reviewers trained in the same way reviewing the same patient record.  To perform inter-rater reliability testing we created an electronic data collection tool and trained the reviewers (raters) on its’ use. The reviewers separately reviewed every sampled patient and collected all data elements necessary for computation of the performance measure (contained on the electronic data collection tool). Data received was analyzed using SAS to calculate frequencies, level of agreement, and agreement statistics (Cohen’s Kappa). Cohen’s Kappa coefficient measures inter-rater agreement for qualitative items and takes into account any agreement occurring by chance.

The following table displays the interpretation of the kappa statistic:

**Kappa Strength of Agreement[[4]](#footnote-4)**  
0.00 Poor  
0.01 - 0.20 Slight  
0.21 - 0.40 Fair  
0.41 - 0.60 Moderate  
0.61 - 0.80 Substantial  
0.81 - 0.99 Almost Perfect

Evidence of content validity is provided by looking for agreement among subject matter experts. The performance measure was assessed for content validity by a panel of technical expert work group members during the development process. This subject matter expert panel had representation from measure methodologists, patient advocacy groups, and clinical specialties. Additional input on the content validity of draft measures is obtained through a 30-day public comment period. All comments received are reviewed by the expert work group and the measures adjusted as needed.

For face validity, an external group of clinical and methodological experts assessed the measure for face validity through an on-line survey. The survey introduction provided the following definition of face validity: Face validity is the extent to which an empirical measurement appears to reflect that which it is supposed to “at face value.” Face validity of an individual measure poses the question of how well the definition and specifications of an individual measure appear to capture the single aspect of care or healthcare quality as intended. Face validity of the measure score as an indicator of quality was systematically assessed as follows:

After the measure was fully specified, the expert panel was asked to rate their agreement with the following statement: “The scores obtained from the measure as specified will accurately differentiate quality across providers”.

Scale 1-5, where 1= Strongly Disagree; 3=Neither Agree nor Disagree; 5=Strongly Agree, N/A = Not Applicable

The face validity panel included 22 panel members from the following organizations:

1. American Geriatric Society
2. American College of Occupational and Environmental Medicine
3. Omada Health
4. University of Chicago
5. Drexel University
6. Ascension St. John Detroit
7. American Society of Addiction Medicine
8. Tufts Medical Center
9. Rush University
10. National Institutes of Health
11. NorthShore University Healthcare
12. Northwestern Medicine
13. Rush University
14. Omada Health
15. Northwestern Medicine
16. Centers for Disease Control and Prevention
17. Emory University
18. Cincinnati Children’s
19. Northwestern Medicine
20. Stony Brook Medicine
21. Advocate Healthcare
22. University of California San Francisco

Regarding feasibility of the data elements, a 2018 feasibility assessment was performed to assess the extent to which the required data are readily available, can be captured without undue burden, and are feasible for implementation within electronic health record systems. Two entities participated in the feasibility assessment for this measure.

• Test Site #1: a multispecialty academic medical center using EPIC EHR

• Test Site #2: a medical center using Matrix Care EHR

For this process, a testing methodology using a Data Element Tool (DET) to assess the availability of the data and the technical feasibility and implementation feasibility of the measures was employed. The DET is an Excel workbook designed to capture information that will determine whether or not each site can feasibly collect the data for the measures. It is structured to collect metadata about each data element necessary to construct each measure stored in the EHR. It will also collect information related to integrity and validity of data collection. Specifically, the DET is designed to capture the following information:

1. Data element information: Whether or not the data element is captured in the EHR, the data source application, primary user interface data location, data type, coding system, unit of measure, frequency of collection, and calculability within the measure context.
2. Measure integrity information: An assessment by the testing site as to what degree the measure, as specified, retains the originally stated intention of the measure.
3. Measure validity information: An assessment by the testing site as to what degree the scores obtained from the measure, as specified, will accurately differentiate quality performance across providers.

The DETs collected responses used to assess technical and implementation feasibility for each measure. Measure technical feasibility was defined as “Can my EHR do this?” and measure implementation feasibility was defined as “Will workflow be used consistently?” The responses were captured in the form of a rating using the following responses:

·         “Feasible. Can do today.”

·         “Feasible with workflow mod/changes to EHR.”

·         “Non-feasible. Unable to do today.”

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

The data element validity testing results including Kappa scores are presented below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **% Agree** | **Kappa** | **N** |
|  |  |  |  |  |
| **Site 1** | **Denominator** | 98 | 0.955 | 100 |
| **Numerator** | 100 | 1.000 | 86 |
|  |  |  |  |  |
| **Site 2** | **Denominator** | 83 | 0.504 | 72 |
| **Exclusions** | 76 | 0.511 | 67 |
| **Numerator** | 100 | \*\* | 23 |

\*\*Kappa scores not calculable with multiple non-responses by raters (i.e., all No/No or all Yes/Yes)

For face validity, the panel rating of the validity statement for the measure were as follows:

N = 22; Mean rating = 4.27 and 86% of respondents either agree or strongly agree that this measure can accurately distinguish good and poor quality.

Frequency Distribution of Ratings

5 (Strongly Agree) – 10

4 (Agree) – 9

3 (Neither Agree nor Disagree) – 2

2 (Disagree) – 1

1 (Strongly Disagree) – 0

X (Not Applicable) – 0

For feasibility, overall, the measures are technically “Feasible. Can do today.” in both EHR systems that tested the measures. The majority of the of the data elements are routinely collected as part of clinical care but additional time and programming resources would be needed to implement the missing elements below:

• Laboratory Test, Performed: 2-H Plasma Glucose During a 75g Oral Glucose Tolerance Test

**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?*)

For this measure, we find moderate to high levels of agreement for all data elements tested. Kappa scores ranged from .50 to 1.0, which is considered moderate to almost perfect.

We found a few instances across the measure where more full and accurate information could be found in the manual abstraction process than through electronic reporting. For example, if height is not measured and weight is measured at an office visit, then BMI may not be calculated and reported in a discrete field. In these cases, BMI can be found only by looking at the chart. Additionally, for exclusions related to the identification of comfort measures, hospice care ambulatory, and palliative care are not always stored in discrete fields, even if these are available. Some information is stored in comments/notes in the EHR.

The results of the validity testing demonstrate that this measure is valid, supported by the results of the data element validity testing, content validity, face validity, and feasibility testing that was conducted.

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

**2b2. EXCLUSIONS ANALYSIS**

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

Each site initially pulled a random sample of patients. Site 1 pulled data that met the initial population criteria for the measure, and applied exclusion criteria to the denominator. The site provided a detailed spreadsheet that included the list of exclusions and reasons for exclusions that met the criteria. Site 2 pulled a random sample from the patient population and tested the exclusion criteria and applied inter-rater reliability testing using Cohen’s Kappa Score.

**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*)

Across the measured entities, there were 67 patients excluded for this measure, with a Kappa score of .51. Performance for this measure was 48%. We would expect a performance score within this range. Because the exclusions for this measure are also widely used in other diabetes-related measures, and are based on evidence-based clinical guidelines, the impact on performance is minimal.

**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*)

With a Kappa score of .51, there is moderate agreement of reliability. The exclusions specified for this measure are consistent with other clinical exclusions that are used for measures with this clinical population. The testing results of the exclusions show moderate agreement for this measure.

Furthermore, the individual clinical exclusions specified in this measure are similar/and closely aligned with several already developed NQF endorsed measurement sets. The data elements for the measure exclusions are as follows: patients with diabetes, pregnancy, hospice care, ambulatory, and palliative care. The following NQF-endorsed measures have those data elements, so we are confident that the measure exclusions are appropriate and statistically demonstrate appropriateness:

* Diabetes Care for People with Serious Mental Illness: Hemoglobin A1c (HbA1c) Control (<8.0%): [http://www.qualityforum.org/QPS/2608](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.qualityforum.org_QPS_2608&d=DwMFAg&c=iqeSLYkBTKTEV8nJYtdW_A&r=Wt5PaJx1APcN56vLfkHCQCppqVZ5RXBGvYMx2jsNGHE&m=R7_Hlhxq7lu8dNVqqQmDYZD1dfWeRL8om_nQBpVZgl8&s=6Yp3vM908eGjnWLoCMlDj6j3A7I8c31EHnrILvZMSA4&e=)
  + Diabetes, Hospice Care Ambulatory
* Comprehensive Diabetes Care: Medical Attention for Nephropathy: [http://www.qualityforum.org/QPS/0062](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.qualityforum.org_QPS_0062&d=DwMFAg&c=iqeSLYkBTKTEV8nJYtdW_A&r=Wt5PaJx1APcN56vLfkHCQCppqVZ5RXBGvYMx2jsNGHE&m=R7_Hlhxq7lu8dNVqqQmDYZD1dfWeRL8om_nQBpVZgl8&s=MesaszBBORSH4oC7rUFQFjPZInG71KUp3t17DYFaC50&e=)
  + Hospice Care Ambulatory
* Diabetes: Foot Exam: [http://www.qualityforum.org/QPS/0056](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.qualityforum.org_QPS_0056&d=DwMFAg&c=iqeSLYkBTKTEV8nJYtdW_A&r=Wt5PaJx1APcN56vLfkHCQCppqVZ5RXBGvYMx2jsNGHE&m=R7_Hlhxq7lu8dNVqqQmDYZD1dfWeRL8om_nQBpVZgl8&s=arSkQEnTxRUwWg4nh9_JvSG2XpHac7ZdUDXMsJIJ7Zo&e=)
  + Diabetes
* Comprehensive Diabetes Care: Eye Exam (retinal) performed: [http://www.qualityforum.org/QPS/0055](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.qualityforum.org_QPS_0055&d=DwMFAg&c=iqeSLYkBTKTEV8nJYtdW_A&r=Wt5PaJx1APcN56vLfkHCQCppqVZ5RXBGvYMx2jsNGHE&m=R7_Hlhxq7lu8dNVqqQmDYZD1dfWeRL8om_nQBpVZgl8&s=5kRlnqqG-lHaQyOqci5AqAhmnsaobiug8e4lhpYhwck&e=)
  + Diabetes
* Weight Assessment and Counseling for Nutrition and Physical Activity for Children/Adolescents (WCC) [http://www.qualityforum.org/QPS/0024](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.qualityforum.org_QPS_0024&d=DwMFAg&c=iqeSLYkBTKTEV8nJYtdW_A&r=Wt5PaJx1APcN56vLfkHCQCppqVZ5RXBGvYMx2jsNGHE&m=R7_Hlhxq7lu8dNVqqQmDYZD1dfWeRL8om_nQBpVZgl8&s=poUL4dziWJRyTv7OPfCGevy7AtfNdd8wXKWoc5AFV7c&e=)
  + Pregnancy, Hospice Care Ambulatory
* Depression remission at 12 months: [http://www.qualityforum.org/QPS/0710e](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.qualityforum.org_QPS_0710e&d=DwMFAg&c=iqeSLYkBTKTEV8nJYtdW_A&r=Wt5PaJx1APcN56vLfkHCQCppqVZ5RXBGvYMx2jsNGHE&m=R7_Hlhxq7lu8dNVqqQmDYZD1dfWeRL8om_nQBpVZgl8&s=4iwB6E0QiFxks27hOTafVIz1qqg023wYmamAjtLfZ8E&e=)
  + Palliative Care

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

**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*** [***2b5***](#section2b5)***.***

**N/A**

**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)**

N/A

**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)*

Differences in performance were not tested, however during testing, performance was calculated with performance rates of 0.292 and 0.483 for sites 1 and 2 respectively.

**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*)

Differences in performance were not tested as this measure was tested for data element validity testing.

|  |  |  |
| --- | --- | --- |
|  |  | |
|  | **Site 1** | **Site 2** |
| **Numerator** | 76 | 2688 |
| **Denominator** | 260 | 5567 |
| **Performance Rate** | 0.292 | 0.483 |

**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?*)  
Differences in performance were not tested

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

**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*)

As part of the reliability testing, sites completed data element tables that assessed for missing elements. All elements that were missing on the sample cases were tracked. Since we found instances across the measures where more full and accurate information could be found in the manual abstraction process than through electronic reporting, this seems to be a consistent issue across all types of measures, not just this particular measure, given the nature of EHR capabilities and limitations.

**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*)

We do not have the number of the overall frequency of missing data

**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*)

N/A

1. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Statistics in Medicine. 1992;11:1511-1519. [↑](#footnote-ref-1)
2. [Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy. 2005; 85(3):257-268. [↑](#footnote-ref-2)
3. Landis, J.R. and Koch, G.G. (1977) “The measurement of observer agreement for categorical data” in Biometrics. Vol. 33. pp 159-74. [↑](#footnote-ref-3)
4. Landis, J.R. and Koch, G.G. (1977) “The measurement of observer agreement for categorical data” in Biometrics. Vol. 33. pp 159-74. [↑](#footnote-ref-4)