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

**Measure Number** (*if previously endorsed*)**:** 2371

**Measure Title**: Annual Monitoring for Patients on Persistent Medications

**IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here:** Click here to enter composite measure #/ title

**Date of Submission**: 1/17/2014

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process |
| Efficiency | Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Not applicable

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

**Sample 1:** 1999-2001

**Sample 2:** 2012

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

**Sample 1:** This measure was tested for item-level reliability in five health plans in the HMO Research Network. Plans included group, staff, network and mixed model managed care structure. Enrollment varied from 250,000 to more than 800,000 members. Geographically, the plans include the Northwest, Mountain, Midwest, East and Southwest.

**Sample 2:** This measure was tested for reliability, empirical validity and meaningful difference in performance using data from Medicare, Medicaid and commercial health plans submitting HEDIS data for measurement year 2012. The plans were nationally representative and included both HMOs and PPOs. The plans varied in size from a minimum of 30 patients eligible for one of the indicators to over 300,000 patients eligible for one of the indicators within a single plan. The number of plans included for analysis are described in the table below.

Number of Health Plans Included in Testing Sample 2

|  |  |  |
| --- | --- | --- |
| **Product Line and Type** | | **N Plans** |
| Commercial | HMO | 212 |
| PPO | 199 |
| Medicaid | HMO | 176 |
| Medicare | HMO | 350 |
| PPO | 152 |

**Face validity testing:** This measure was systematically evaluated for face validity with four panels of experts. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliation of expert panel.

* The Medication Management Technical Expert Sub-committee included three physicians and three pharmacists with expertise in medication management.
* The Geriatric MAP included 13 experts in geriatrics, including representation by consumers, health plans, health care providers and policy makers.
* The Technical Measurement Advisory Panel includes 14 members, including representation by health plans methodologists, clinicians and HEDIS auditors.
* NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 21 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

**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*)

**Sample 1**: Medical records were abstracted for a total of 634 patients across the five participating plans based on a random sample of members who received the drugs of interest and who either did or did not receive the laboratory tests of interest. Medical record data was reported consistently across all five field test participant plans. The numbers of medical records reviewed were proportionate to the number of patients of each drug was dispensed to at each site. This resulted in a total of 47 for ARB and ARB combinations, 57 for digoxin, 262 for ACEI and ACEI combinations, and 268 for diuretics and diuretic combinations. The average age across patients was 63 years old and 50% were female.

**Sample 2**: In 2012, HEDIS measures covered 99.4 million commercial health plan beneficiaries, 14.3 million Medicaid beneficiaries, and 11.5 million Medicare beneficiaries. Data is summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid) and type (HMO, PPO). The patients included in the HEDIS data include a diverse representation of ages, race and diagnoses. The table below shows the average number of eligible patients per health plan and the standard deviation of that average across health plans.

Average Eligible Population (EP) per Health Plan Included in Testing Sample 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **N Plans** | **Avg. EP** | **SD** |
| Commercial | HMO | 212 | 17196.1 | 37726.9 |
| PPO | 199 | 23707.5 | 41368.6 |
| Medicaid | HMO | 176 | 5449.1 | 6722.9 |
| Medicare | HMO | 350 | 16397.4 | 32151.6 |
| PPO | 152 | 12553.0 | 22672.6 |

EP: Eligible population

SD: Standard deviation

**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**.

Sample 1 was used to test item-level validity.

Sample 2 was used to test reliability, empirical validity, and meaningful difference in performance

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

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability Testing of Performance Measure Score: In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009). The following is quoted from the tutorial which focused on provider-level assessment: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it 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. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.” This approach is also relevant to health plans and other accountable entities.

Adams’ approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

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*)

Results of Reliability Testing of Performance Measure Score:

Beta-Binomial Statistic For Each Measure Rate

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rate | **Commercial** | | | **Medicaid** | | | **Medicare** | | |
| Avg. | SD | 10-90th | Avg. | SD | 10-90th | Avg. | SD | 10-90th |
| **ACE/ ARBs** | 0.96 | 0.06 | 0.89-0.99 | 0.90 | 0.15 | 0.73-0.99 | 0.95 | 0.10 | 0.86-1.00 |
| **Digoxin** | 0.72 | 0.17 | 0.46-0.94 | 0.36 | 0.17 | 0.13-0.59 | 0.71 | 0.23 | 0.35-0.96 |
| **Diuretics** | 0.95 | 0.08 | 0.85-1.00 | 0.89 | 0.15 | 0.71-0.99 | 0.93 | 0.12 | 0.81-0.91 |
| **Total** | 0.81 | 0.04 | 0.75-0.86 | 0.86 | 0.05 | 0.80-0.91 | 0.92 | 0.04 | 0.87-0.96 |

Avg.: Average beta-binomial

SD: Standard deviation across plans

10-90th: 10th percentile-90th percentile across plans

**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?*)

Reliability Testing of Performance Measure Score: Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (signal). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The testing suggests the all three indicators within this measure have good reliability between 0.7 and 1.0. The only exception is Digoxin in the Medicaid population, the average reliability is 0.36. This low reliability is due primarily to small sample size for this measure in 2012. Less than half of Medicaid plans were able to meet the minimum sample size of 30 patients with a prescription for Digoxin necessary to report this measure. Of those plans that did report this measure, the average denominator size was much lower than the average denominator size for Medicare and commercial health plans. This is due to the small size of Medicaid plans in general. Reliability is not a static property of a measures and as the number of Medicaid HMOs continues to grow we expect the reliability of the measure score to increase for this one indicator.

The 10-90th percentile distribution of health plan level-reliability on the rates in this measure show the vast majority of health plans met or exceeded the minimally accepted threshold of 0.7, and the majority of plans exceeded 0.8. Strong reliability is demonstrated since majority of variances is due to signal and not to noise.

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

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*  
Method of testing critical data element validity: We tested critical data element validity by comparing medical records (considered the “gold standard”) with administrative data. We evaluated the accuracy of the administrative data to determine the sensitivity, specificity and positive predictive value (PPV) of the three indicators at each health plan. The sensitivity is the proportion of patients who had the laboratory test conducted that was documented in the medical record who also have an administrative claims code for the laboratory test (i.e., the true positives divided by the sum of the true positives plus the false negatives). Specificity is the proportion of patients who did not have the laboratory test conducted who also did not have the claims code for the laboratory test (i.e., the true negatives divided by the sum of the true negatives plus the false positives). The PPV is the proportion of patients who had the laboratory test conducted that are true positives (i.e., the true positives divided by the true positives plus the false positives).

Method of empirical validity testing: We tested for construct validity by exploring whether the indicators within this measure were correlated with each other. We hypothesized that organizations that perform well on one of the three indicators should perform well on the measures. To test these correlations we used a Person correlation test. The Pearson correlation estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +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.

Systematic Assessment of Face Validity: NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Measurement Advisory Panels (MAPs) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s MAPs, the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.   
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

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

STEP 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA’s Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

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

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
Across all sites and all ages combined, the sensitivity, specificity, and PPV by drug and by laboratory test is shown below.

Results of testing critical data element validity:

|  |  |  |  |
| --- | --- | --- | --- |
| ***Drug*** | ***Sensitivity (%)*** | ***Specificity (%)*** | ***PPV (%)*** |
| ACEI and ACEI Combinations  Serum creatinine  Serum potassium | 97  97 | 83  87 | 95  95 |
| ARB and ARB Combinations  Serum creatinine  Serum potassium | 89  89 | 67  82 | 92  94 |
| Digoxin  Serum creatinine  Serum potassium | 93  98 | 62  69 | 89  89 |
| Diuretics and Diuretic Combinations  Serum creatinine  Serum potassium | 96  96 | 91  80 | 98  95 |

\*Serum Digoxin was added to the measure after initial field testing, therefore we do not have critical data element validity data for this monitoring parameter.

Results of empirical validity testing:

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACE inhibitors or ARBs | Digoxin | Diuretics |
| ACE inhibitors or ARBs | - | R=0.76519 | R=0.98173 |
| - | p<.0001 | p<.0001 |
| Digoxin | - | - | R=0.76525 |
| - | - | p<.0001 |

Results of Systematic Assessment of Face Validity

Step 1: This measure was developed to address medication safety for commonly used long-term medications in 2004. NCQA and the Geriatric MAP worked together to assess the most appropriate tools for evaluating health plan performance at monitoring individuals taking common persistent medications.

Step 2: The measure was written and field-tested in 2004. After reviewing field test results, the CPM recommended to send the measure to public comment with a majority vote in 2005.

Step 3: The measure was released for Public Comment in 2005 prior to publication in HEDIS. We received and responded to 32 comments on this measure. The CPM recommended moving this measure to first year data collection by a majority vote.

Step 4: The measure was introduced in HEDIS 2006. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure public reporting with a majority vote.

Step 5: The measure is currently undergoing re-evaluation

Conclusion: The measure was deemed to have the desirable attributes of a HEDIS measure in 2005 (relevance, scientific soundness, and feasibility).

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Interpretation of testing critical data element validity: These results suggest that administrative data at health plans is sufficiently accurate to use as a proxy for evaluating whether laboratory tests are completed. The presence of a claims code for the laboratory tests including potassium and serum creatinine had good to excellent (89 to 99%) sensitivity and PPV (89 to 98%). Specificity of the absence of a claims code for one of these laboratory tests is also very good (62 to 91%). Serum Digoxin was added to the measure as a monitoring parameter in 2013, therefore it was not included in the initial field test. However, we assume the sensitivity and specificity of administrative claims for detecting laboratory tests is consistent across the specific type of test and can be extrapolated to serum digoxin.

Interpretation of empirical validity testing: 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. The results confirmed the hypothesis that the three monitoring indicators are correlated very highly with each other (0.8-1.0), suggesting they represent the same underlying quality construct of persistent medication monitoring quality. These correlation coefficients were significantly different from 0 (p<0.001).

Interpretation of systematic assessment of face validity: These results indicate the technical expert panel showed good agreement that the measures as specified will accurately differentiate quality across providers. Our interpretation of these results is that this measure has sufficient face validity.

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

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)  
This measure includes an exclusion for individuals who were admitted to an inpatient facility in the measurement year. This exclusion was included because it is assumed any individual admitted to an inpatient facility will be given a standard lab panel that includes the monitoring parameters included in this measure but may not be separately represented as individual claims codes (i.e. they may be included in a DRG or other bundled payment).

In HEDIS data submission, the process of identifying individuals eligible for exclusions is audited by independent HEDIS certified auditors but is not reported. Therefore we cannot provide data on the number of exclusions in this measure.

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)  
None

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)  
None

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

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** 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

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

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

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

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

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

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

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

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

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

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

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

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. 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 .05, then the two plans’ performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However the method can be used for comparison of any two measured entities.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
**Variation in Performance across Health Plans: Measurement Year 2012**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Rate** | **Avg. EP** | **Avg.** | **SD** | **10th** | **25th** | **50th** | **75th** | **90th** | **IQR** | **p-value** |
| **Commercial** | **ACE/ARB** | 11879 | 81.1 | 4.4 | 75.9 | 78.6 | 81.2 | 84.0 | 86.6 | 5.4 | <0.001 |
| **Digoxin** | 206 | 83.3 | 6.9 | 74.3 | 79.3 | 83.9 | 87.8 | 91.9 | 8.5 | <0.001 |
| **Diuretics** | 7978 | 80.7 | 4.6 | 75.2 | 78.1 | 80.7 | 83.5 | 86.2 | 5.4 | 0.01 |
| **Total** | 19893 | 81.0 | 4.4 | 75.8 | 78.5 | 81.0 | 83.8 | 86.5 | 5.3 | <0.001 |
| **Medicaid** | **ACE/ARB** | 2813 | 86.3 | 4.7 | 80.7 | 84.6 | 87.0 | 89.2 | 91.2 | 4.6 | <0.001 |
| **Digoxin** | 106 | 90.1 | 4.3 | 83.7 | 87.4 | 90.7 | 93.2 | 94.9 | 5.8 | <0.001 |
| **Diuretics** | 2093 | 85.9 | 5.1 | 79.0 | 83.7 | 86.7 | 89.1 | 91.3 | 5.4 | <0.001 |
| **Total** | 4798 | 86.1 | 4.7 | 80.4 | 84.0 | 87.0 | 89.2 | 91.1 | 4.9 | <0.001 |
| **Medicare** | **ACE/ARB** | 8412 | 91.9 | 4.5 | 88.4 | 90.7 | 92.3 | 94.0 | 95.8 | 3.4 | <0.001 |
| **Digoxin** | 500 | 94.1 | 3.5 | 90.2 | 92.7 | 94.4 | 96.1 | 97.8 | 3.4 | <0.001 |
| **Diuretics** | 6173 | 92.1 | 4.6 | 88.5 | 90.9 | 92.6 | 94.2 | 96.0 | 3.3 | <0.001 |
| **Total** | 14878 | 92.0 | 4.5 | 88.6 | 90.8 | 92.5 | 94.1 | 95.9 | 3.4 | <0.001 |

EP: Eligible Population, the average denominator size across plans submitting to HEDIS

IQR: Interquartile range

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

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)  
The results above indicate there is a 3-9% gap in performance between the 25th and 75th performing plans. For all product lines and rates the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for commercial which show a 5.3-8.5 percentage point gap between 25th and 75th percentile plans. This gap represents on average 641 more ACE/ARB patients, 17 more digoxin patients and 431 more diuretic patients receiving monitoring in high performing commercial HMO plans compared to low performing plans (estimated from average health plan eligible population).

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

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

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

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

This measure is collected using all available administrative claims, there is no missing data on this measure.

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)  
This measure is collected using all available administrative claims, there is no missing data on this measure.

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

This measure is collected using all available administrative claims, there is no missing data on this measure.