**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**: INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications

**Date of Submission**: 1/16/2014

**Type of Measure: Process**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | ☐ Outcome (*including PRO-PM*) |
| ☐ Cost/resource | X 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.  **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.    **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**.  **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 |
| X administrative claims | X 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*).

The following specific datasets were used for testing:

* 2007-2008 Medicare Parts A, B, and D claims data for eight states (Arizona, Delaware, Florida, Indiana, Iowa, Mississippi, Rhode Island, and Washington)
* 2011-2012 Medicare Parts A, B, and D claims data for 10 states (Arizona, Delaware, Florida, Indiana, Iowa, Mississippi, Missouri, Rhode Island, Texas, and Washington)
* 2011 Medicare Parts A, B, and D claims data for 31 ACOs

**1.3. What are the dates of the data used in testing**? January 1, 2007 – December 31, 2008; January 1, 2011 – December 31, 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 |
| X health plan | X health plan |
| X other: State, Accountable Care Organization | X other: State, Accountable Care Organization |

**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*)   
Characteristics of the sample for 2007-2008 are summarized in Table 1. All beneficiaries from eight states (Arizona, Delaware, Florida, Indiana, Iowa, Mississippi, Rhode Island, and Washington) were included in the testing sample. Measured entities included eight states, 93 prescription drug plans (PDPs), and 13,023 physician groups. Fifteen percent of PDPs had fewer than 30 beneficiaries attributed, accounting for less than 0.01% of total beneficiaries attributed to a prescription drug plan. Likewise, 71% of physician groups had less than 30 beneficiaries attributed; they represent only 4.0% of the total number of beneficiaries attributed to a physician group.

**Table 1. 2007-2008 Sample Characteristics by PDPs and Physician Groups**

|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristics** | **States**  **n=8** | **Prescription Drug Plans n=93** | **Physician Groups n=13,023** |
| Total Number | 4,789,034 | 4,789,034 | 4,789,034 |
| Total Attributed (%) | 4,789,034 (100%) | 2,438,280 (50.91%) | 1,113,012 (23.24%) |
| Mean # of Beneficiaries | 598,629 | 0,026,218 | 90 |
| Median # of Beneficiaries | 444,929 | 0,000555 | 4 |
| Min # of Beneficiaries | 79,875 | 0,000001 | 1 |
| Max # of Beneficiaries | 2,146,482 | 0,474,068 | 10,950 |
| STD | 655,036 | 0,075,401 | 367 |
| P10 | 79,875 | 0,000,015 | 1 |
| P25 | 242,068 | 0,000,092 | 1 |
| P50 | 444,929 | 0,000,555 | 4 |
| P75 | 594,343 | 0,009,855 | 46 |
| P90 | 2,146,482 | 0,074,360 | 190 |

Characteristics of the sample for 2011-2012 are summarized in Table 2. All beneficiaries from 10 states (Arizona, Delaware, Florida, Indiana, Iowa, Mississippi, Missouri, Rhode Island, Texas, and Washington) were included in the testing sample. Measured entities included 10 states, 83 prescription drug plans (PDPs), and 26,182 physician groups. Fourteen percent of PDPs had fewer than 30 beneficiaries attributed, accounting for less than 0.01% of total beneficiaries attributed to a prescription drug plan. Sixty-five percent of physician groups had fewer than 30 beneficiaries attributed. These groups represent 1.2% of the total number of beneficiaries attributed to a physician group.

**Table 2. 2011-2012 Sample Characteristics by States and Physician Groups**

|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristics** | **States**  **n=10** | **Prescription Drug Plans n=83** | **Physician Groups n=26,182** |
| Total Number | 14,162,440 | 14,162,440 | 14,162,440 |
| Total Attributed (%) | 14,162,440 (100%) | 4,699,420 (33.18%) | 4,241,116 (29.95%) |
| Mean # of Beneficiaries | 1,416,244 | 56,656 | 194 |
| Median # of Beneficiaries | 1,171,694 | 1,221 | 10 |
| Min # of Beneficiaries | 183,084 | 1 | 1 |
| Max # of Beneficiaries | 4,098,325 | 1,102,813 | 37,977 |
| STD | 1,369,273 | 167,654 | 907 |
| P10 | 200,154 | 8 | 1 |
| P25 | 598,022 | 113 | 3 |
| P50 | 1,171,694 | 1,221 | 10 |
| P75 | 1,213,975 | 38,693 | 85 |
| P90 | 3,896,824 | 121,506 | 394 |

A convenience sample of beneficiaries attributed to 31 Accountable Care Organizations (ACOs) was used for testing the measure at the ACO level. Characteristics of the ACO sample for 2011 are summarized in Table 3.

**Table 3. 2011 Sample Characteristics for 31 ACOs**

| **Characteristics** | **ACOs** |
| --- | --- |
| Total Number | 31 |
| Total Beneficiaries | 682,036 |
| Mean # of Beneficiaries | 22,001 |
| Median # of Beneficiaries | 18,622 |
| Min # of Beneficiaries | 7,207 |
| Max # of Beneficiaries | 61,957 |
| STD | 12,001 |
| P10 | 10,309 |
| P25 | 13,249 |
| P50 | 18,622 |
| P75 | 24,356 |
| P90 | 35,853 |

**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*)

Demographic characteristics of the beneficiaries in the 2007-2008 and 2011-2012 datasets are shown in Tables 4 and 5, respectively.

**Table 4. 2007-2008 Demographic Characteristics by States, PDPs, and Physician Groups**

| **Characteristics** | **States**  **n=8** | **Prescription Drug Plans n=93** | **Physician Groups n=13,023** |
| --- | --- | --- | --- |
| **Total Population** | 4,789,034 | 2,438,280 | 1,113,012 |
| **Gender** | | | |
| Female | 2,802,453 (58.52%) | 1,512,475 (62.03%) | 710,962 (63.88%) |
| Male | 1,986,202 (41.47%) | 925,805 (37.97%) | 402, 050 (36.12%) |
| Unknown | 379 (0.01%) | Not applicable | Not applicable |
| **Age** | | | |
| ≥65 years | 3,556,636 (74.27%) | 1,827,717 (74.96%) | 897,883 (80.67%) |
| **Race** | | | |
| White/Caucasian | 3,959,024 (82.67%) | 2,031,546 (83.32%) | 967,361 (86.91%) |
| African-American | 487,219 (10.17%) | 247,149 (10.14%) | 88,129 (7.92%) |
| Hispanic | 185,131 (3.87%) | 86,628 (3.55%) | 29,888 (2.69%) |
| Other | 157,660 (3.29%) | 72,957 (2.99%) | 27,634 (2.48%) |
| **Ethnicity** | | | |
| Hispanic | 185,131 (3.87%) | 86,628 (3.55%) | 29,888 (2.69%) |
| Non-Hispanic | 4,603,903 (96.13%) | 2,351,652 (96.45%) | 1,083,124 (97.31%) |
| **Medicare and Medicaid Eligibility** | | | |
| Dual Eligible | 1,216,287 (25.40%) | 810,110 (33.22%) | 308,649 (27.73%) |
| Non-Dual Eligible | 3,572,747 (74.60%) | 1,628,170 (66.78%) | 804,363 (72.27%) |

**Table 5. 2011-2012 Demographic Characteristics by State, PDPs, and Physician Groups**

| **Characteristics** | **State**  **n=10** | **Prescription Drug Plans n=83** | **Physician Groups n=26,182** |
| --- | --- | --- | --- |
| **Total Population** | 14,162,440 | 4,699,420 | 4,241,116 |
| **Gender** | | | |
| Female | 6,948,546 (49.06%) | 2,697,239 (57.40%) | 2,482,734 (58.54%) |
| Male | 5,827,374 (41.15%) | 1,782,594 (37.93%) | 1,710,539 (40.33%) |
| Unknown | 1,386,520 (9.79%) | 219,587 (4.67%) | 47,843 (1.13%) |
| **Age** | | | |
| ≥65 years | 9,949,181 (70.25%) | 3,326,257 (70.78%) | 3,334,085 (78.61%) |
| **Race** |  |  |  |
| White/Caucasian | 11,086,802 (78.28%) | 3,887,785 (82.73%) | 3,693,852 (87.10%) |
| African-American | 1,213,508 (8.57%) | 460,400 (9.80%) | 335,859 (7.92%) |
| Hispanic | 474,632 (3.35%) | 195,928 (4.17%) | 109,142 (2.57%) |
| Other | 1,387,498 (9.80%) | 155,307 (3.30%) | 89,041 (2.10%) |
| **Ethnicity** | | | |
| Hispanic | 474,632 (3.35%) | 195,928 (4.17%) | 109,142 (2.57%) |
| Non-Hispanic | 13,687,808 (96.65%) | 4,503,492 (95.83%) | 4,131,974 (97.43%) |
| **Medicare and Medicaid Eligibility** | | | |
| Dual Eligible | 2,029,697 (14.33%) | 1,339,687 (28.51%) | 785,130 (18.51%) |
| Non-Dual Eligible | 12,132,743 (85.67%) | 3,359,733 (71.49%) | 3,455,986 (81.49%) |

Demographic characteristics of the beneficiaries in the ACO dataset are shown in Table 6.

**Table 6. 2011 Demographic Characteristics by ACO**

| **Characteristics** | **ACO Number (%)** |
| --- | --- |
| **Total Population** | 682,036 |
| **Gender** | |
| Female | 398,763 (58.47%) |
| Male | 283,273 (41.53%) |
| **Age** | |
| ≥65 years | 574,224 (84.19%) |
| **Race** | |
| White/Caucasian | 574,672 (84.26%) |
| African-American | 46,211 (6.78%) |
| Hispanic | 21,310 (3.12%) |
| Other | 38,181 (5.60%) |
| Unknown | 1,662 (0.24%) |
| **Ethnicity** | |
| Hispanic | 21,310 (3.12%) |
| Non-Hispanic | 660,726 (96.88%) |
| **Medicare and Medicaid Eligibility** | |
| Dual Eligible | 152,960 (22.43%) |
| Non-Dual Eligible | 529,076 (77.57%) |

**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**.

Not applicable

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

**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*)   
X **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*)

The method of reliability testing used and the rationale are described below.

Method of Reliability Testing and Rationale

In order to assess measure precision in the context of the observed variability across measurement units (states, prescription drug plans [serving as a proxy for health plans], and Accountable Care Organizations [ACOs]), we utilized the approach proposed by Adams (2009) and Scholle et al. (2008). The rationale for this choice of testing was based upon the work on the reliability of provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial published by Adams: “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.”

The signal-to-noise ratio was calculated as a function of the variance between measured entities (signal) and the variance within a measured entity (noise). Reliability was estimated using a beta-binomial model. This approach has two basic assumptions:

1. Each measured entity has a true pass rate, p, which varies from group to group; and,
2. The measured entity’s score is a binomial random variable conditional on the entities true value, which comes from the beta distribution.

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 physician group variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across physician groups). In a simulation, Adams showed that differences between physicians started to be seen at reliability of 0.7 and significant differences could be seen at reliability of 0.9. Our rationale was based on Adams’ work, and thus, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to discriminate performance between physicians.

# Using methodology described by Scholle et al. (2008), reliability estimates were computed separately based on the mean denominator size for physicians within each denominator category. As Scholle described in the article, the reliability estimate at the mean denominator for each category should reflect “the typical experience of physicians in this population.”

Reliability scores were also calculated for state, prescription drug plan (which served as a proxy for health plans), and ACO levels of measurement using the same approach.

Citations

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

Scholle, S. H., Roski, J., Adams, J. L., Dunn, D. L., Kerr, E. A., Dugan, D. P., et al. (2008). Benchmarking physician performance: Reliability of individual and composite measures. *American Journal of Managed Care, 14*(12), 833-838.

**2a2.3. For each level 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*)

State Level

We conducted reliability tests across measurement units, and the results from the state level, including reliability statistics and assessments of adequacy, are stated below (Table 7).

**Table 7.** **2011-2012 State Reliability and Assessment of Adequacy for Tests Conducted**

| **State** | **Measure Rate (Reliability)** |
| --- | --- |
| AZ | 19.92% (0.98) |
| DE | 22.68% (0.92) |
| FL | 21.54% (0.99) |
| IA | 23.33% (0.99) |
| IN | 21.51% (0.99) |
| MO | 21.18% (0.99) |
| MS | 16.09% (0.99) |
| RI | 32.04% (0.88) |
| TX | 18.20% (0.99) |
| WA | 23.36% (0.98) |

Prescription Drug Plan Level

Using the method of mean denominator and volume categories, a minimum denominator of 2,500 resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between plans. The aforementioned criteria resulted in 14.5% of all plans (10 of 69 Plans) with reliable scores (Table 8).

**Table 8. 2012 Prescription Drug Plan Reliability and Assessment of Adequacy for Tests Conducted**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Min Denominator** | **# of Plans (% of PDPs with at least 1 individual attributed) n=69** | **Mean Rate of Plans** | **Reliability Score** |
| 2012 | 2,500 | 10 (14.49%) | 20.69% | 0.71 |

ACO Level

Using the method of mean denominator and volume categories, a minimum denominator of 180 resulted in an overall reliability score of 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between ACOs. The aforementioned criteria resulted in 77.4% of all ACOs (24 of 31 ACOs) with reliable scores (Tables 9 and 10).

**Table 9. 2011 Overall ACO Reliability and Assessment of Adequacy for Tests Conducted**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Min Denominator** | **# of ACOs (% of ACOs with at least 1 individual attributed)** | **Mean Rate of ACOs** | **Reliability Score** |
| 2011 | 180 | 24 (77.4%) | 22.62% | 0.70 |

**Table 10. 2011 Individual ACO Reliability and Assessment of Adequacy for Tests Conducted**

| **ACO #** | **Denominator** | **Measure Rate (Reliability)** |
| --- | --- | --- |
| 1 | 423 | 32.86% (0.81) |
| 2 | 228 | 26.32% (0.73) |
| 3 | 423 | 23.17% (0.84) |
| 4 | 210 | 15.24% (0.77) |
| 5 | 202 | 15.84% (0.76) |
| 6 | 236 | 23.73% (0.75) |
| 7 | 174 | 13.22% (0.74) |
| 8 | 288 | 23.96% (0.78) |
| 9 | 315 | 18.10% (0.82) |
| 10 | 181 | 19.89% (0.71) |
| 11 | 527 | 22.77% (0.87) |
| 12 | 161 | 23.60% (0.67) |
| 13 | 442 | 26.24% (0.84) |
| 14 | 313 | 21.41% (0.81) |
| 15 | 377 | 20.95% (0.83) |
| 16 | 539 | 26.16% (0.86) |
| 17 | 381 | 29.66% (0.81) |
| 18 | 179 | 25.14% (0.69) |
| 19 | 216 | 14.81% (0.78) |
| 20 | 260 | 30.38% (0.74) |
| 21 | 269 | 24.54% (0.77) |
| 22 | 36 | 16.67% (0.33) |
| 23 | 620 | 25.48% (0.88) |
| 24 | 331 | 17.52% (0.83) |
| 25 | 165 | 20.00% (0.69) |
| 26 | 144 | 17.36% (0.68) |
| 27 | 168 | 13.69% (0.73) |
| 28 | 185 | 20.00% (0.72) |
| 29 | 207 | 13.04% (0.78) |
| 30 | 270 | 32.22% (0.74) |
| 31 | 874 | 18.54% (0.93) |

**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 results indicate that the measure as currently specified was reliable at the state, prescription drug plan, and ACO levels.  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**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*)

X **Performance measure score**

X **Empirical validity testing**X **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 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)*

Performance Measure Score

1. Convergent Validity - Relationship to another measure as expected (NQF 0556 compared to NQF 0555), Pearson Correlation Score
2. Threats to Validity, Analysis of Missing Data, Frequency
3. ICD10-CM Conversion Methodology

Convergent Validity

We compared a related NQF-endorsed measure, NQF 0555, which assesses INR monitoring for individuals on warfarin. We would expect the scores on these measures to be correlated since they reflect a similar concept of timely and appropriate INR monitoring. We tested the measure distributions for normality at each unit of analysis and then selected the appropriate statistical test for the distribution and assessed the significance of the correlation coefficient.

Face Validity Method

FMQAI’s Technical Expert Panel (TEP) evaluated the face validity of the measure and measure score after field testing was completed. The names and organizations of TEP members are listed in Table 11. The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®). TEP members were specifically asked whether “the performance score from the measure as specified represents an accurate reflection of quality of care.” They responded by indicating their level of agreement with the statement on a 5-point Likert scale (1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree).

**Table 11. TEP Members**

| **Name** | **Organization** |
| --- | --- |
| **Dale W. Bratzler**  DO, MPH | Professor and Associate Dean, College of Public Health, University of Oklahoma Health Sciences Center |
| **Mary Brennan-Taylor** | Adjunct Research Instructor of Family Medicine, School of Medicine and Biomedical Sciences, University of Buffalo  Representing: TEP as Patient Representative |
| **Frank E. Briggs III** PharmD, MPH | Vice President, Quality and Patient Safety, West Virginia University Healthcare  Representing: American Society of Health-System Pharmacists |
| **Daniel Castillo**  MD, MBA | Medical Director, Healthcare Quality Evaluation, The Joint Commission |
| **Joan Ching**  RN, MN, CPHQ | Administrative Director, Hospital Quality & Safety, Virginia Mason Medical Center |
| **Edward S. Eisenberg**  MD, FACP | Senior Vice President, Performance Measurement and Strategic Alliances, Pharmacy Quality Alliance |
| **Floyd Eisenberg**  MD, MPH, FACP | President, iParsimony, LLC |
| **Marybeth Farquhar**  PhD, MSN, RN | Vice President of Research & Measurement, URAC |
| **Frank Federico**  BS, RPh | Executive Director for Strategic Partners, Institute for Healthcare Improvement |
| **Robert Feroli**  PharmD, FASHP | Medication Safety Officer, Johns Hopkins Hospital |
| **Tejal Gandhi**  MD, MPH | President, National Patient Safety Foundation; Board-certified Internist and Associate Professor of Medicine, Harvard Medical School  Representing: American Hospital Association |
| **P. Michael Ho**  MD, PhD, FACC | Staff Cardiologist, VA Eastern Colorado Health Care System; Associate Professor of Medicine, University of Colorado Denver  Representing: American College of Cardiology |
| **Mark L. Holtsman** PharmD | Co-Director, Inpatient Pain Service and Pain Management Service Pharmacist, UC Davis Medical Center; Clinical Professor of Anesthesiology and Pain Medicine, UC Davis School of Medicine  Representing: American Academy of Pain Medicine |
| **Clifford Ko**  MD, MS, MSHS, FACS | Director, ACS Division of Research and Optimal Patient Care; Director, ACS NSQIP; Professor of Surgery and Health Services, UCLA Schools of Medicine and Public Health  Representing: American College of Surgeons |
| **Janet Maurer**  MD, MBA, FCCP | Operations Medical Director, National Imaging Associates, Health Dialog; Clinical Professor of Medicine, University of Arizona, College of Medicine, Phoenix Campus; Staff Physician, St. Joseph’s Medical Center  Representing: American College of Chest Physicians |
| **Michael N. Neuss**  MD | Chief Medical Officer, Vanderbilt-Ingram Cancer Center; Professor of Medicine, Vanderbilt School of Medicine  Representing: American Society of Clinical Oncology |
| **N. Lee Rucker**  MSPH | Senior Advisor, National Council on Patient Information and Education |
| **Edward Septimus**  MD, FACP, FIDSA, FSHEA | Medical Director, Infection Prevention and Epidemiology Clinical Service Group, HCA Healthcare System; Clinical Professor of Internal Medicine, Texas A & M University  Representing: Infectious Diseases Society of America |
| **Nathan Spell**  MD, FACP | Chief Quality Officer, Emory University Hospital; Associate Professor of Medicine, Emory University School of Medicine  Representing: American College of Physicians |
| **Stephen J. Traub**  MD, FACEP | Assistant Professor in Emergency Medicine and Chair, Department of Emergency Medicine, Mayo Clinic  Representing: American College of Emergency Physicians |
| **Darren M. Triller**  PharmD | Senior Director, Quality Improvement, IPRO QIO |

Threats to Validity

Days’ supply is a critical variable in determining warfarin usage. We assessed all warfarin claims for patients in the denominator for missing days’ supply. Specifically, for missing days’ supply, we analyzed the number (%) of beneficiaries in the measure denominator with one or more claims that had missing days’ supply.

ICD-10-CM Conversion Methodology

The conversion of the measure to include ICD-10-CM codes is provided as requested by NQF. The crosswalk is provided as an Excel file in Section 2a1.3, Data Dictionary or Code Table.

**Statement of Intent**

The goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

**Name and Credentials of Experts Who Assisted in the Process**

* Soeren Mattke, MD, DSc, Senior Scientist, RAND Corporation
* Tim Laios, MBA, MPH, Executive Director, Informatics, Health Services Advisory Group (HSAG)
* Ryan Fair, BS, Director, Informatics, HSAG
* Kerri Carlile, MS, Informatics Analyst, HSAG
* Sara Lomeli, BA, Informatics Project Coordinator, HSAG

**Methodology – Evaluation of ICD-9-CM Changes**

The changes (i.e., deletions and/or additions) made to the ICD-9-CM codes for the measures requiring conversion were reviewed. Additionally, the ICD-9-CM codes were reviewed for any coding updates, using the October 2013 Conversion Table of New ICD-9-CM Codes, published by the National Center for Health Statistics (NCHS) and the Centers for Medicare & Medicaid Services (CMS).

**ICD-9-CM Code Identification**

For each measure requiring conversion, original tables were used to identify all ICD-9-CM codes included in the measure. Those ICD-9-CM codes and matching descriptions were then extracted from the Ingenix 2011 ICD-9-CM data file. Only valid ICD-9-CM codes were retained and used in the ICD-9-CM to ICD-10-CM conversion process.

**Ingenix Extraction**

When extracting the ICD-9-CM codes from the Ingenix data file, all codes were extracted with two-decimal specificity. For example, for ICD-9-CM code 274.1, all ICD-9-CM codes that had 2741 for the first four digits were extracted (e.g., 274.10, 274.11, and 274.19). For every three-digit ICD-9-CM code used in the measure, all ICD-9-CM codes that began with those first three digits were extracted. For the ICD-9-CM codes listed in ranges, codes with up to two-decimal specificity were extracted within that range.

**Conversion Process**

The ICD-9-CM and ICD-10-CM General Equivalence Map (GEM) text files and the ICD-10-CM Descriptions text file were imported into SAS[[1]](#footnote-2) and combined into one data file to capture all ICD-9-CM codes, their corresponding ICD-10-CM codes, and the ICD-10-CM code descriptions.[[2]](#footnote-3),[[3]](#footnote-4) The ICD-9-CM codes that were retained from the Ingenix 2011 ICD-9-CM Data File described above were then extracted from the combined GEM data file.

The results for each measure were then exported into Microsoft Excel[[4]](#footnote-5) and validated to ensure that the translation was appropriate (i.e., the crosswalk was correct and applied appropriately and all appropriate ICD-9-CM codes were captured). Since one ICD-9-CM code can have several corresponding ICD-10-CM codes, each ICD-9-CM code can have multiple entries in the final Excel document (i.e., one row for each corresponding ICD-10-CM code).

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

Convergent Validity

The measure rate is positively correlated with the NQF-endorsed measure, INR Monitoring for Individuals on Warfarin (NQF #0555), at the ACO level (ρ=0.745, p-value<0.0001). The distribution of the measure rates is presented in Table 12.

**Table 12. Distribution of Measure Rates – ACO**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Measure** | **Count ACO** | **Mean Measure Rate** | **Standard Deviation** | **Median** | **Minimum** | **Maximum** |
| INR Monitoring for Individuals on Warfarin (NQF #0555) | 31 | 75.3% | 9.8% | 78.0% | 58.5% | 90.7% |
| INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications (NQF #0556) | 31 | 21.7% | 5.5% | 21.4% | 13.0% | 32.9% |

Systematic Assessment of Face Validity

Sixteen of the 21 (76.2 %) TEP members completed the face validity evaluation for the measure. The results of the TEP rating of face validity on a scale of 1 to 5 are presented in Table 13.

**Table 13. Results of the Face Validity Evaluation**

|  |  |
| --- | --- |
| **Rating** | **Number of TEP (%)** |
| 5 (Strongly Agree) | 6 (37.5%) |
| 4 (Agree) | 9 (56.3%) |
| 3 (Neutral) | 1 (6.3%) |
| 2 (Disagree) | 0 |
| 1 (Strongly Disagree) | 0 |

Of the TEP members who evaluated the measure for face validity, 93.8% (15/16) strongly agreed or agreed that the measure was valid as specified. The mean rate was 4.3, and the median rate was 4.

Threats to Validity

Percentage of individuals in the denominator with one or more claims with missing days’ supply (0%)

**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?*)

Convergent Validity

The measure rates between NQF 555 and NQF 556 were strongly correlated (>0.7) as expected, and this adds further support that the measure as specified is valid.

Face Validity

In summary, 93.8% of TEP members who responded to the survey strongly agreed or agreed that the measure has face validity.

Threats to Validity

All claims in the analysis had the days’ supply field populated. Therefore, no impact on the accuracy of the measure is expected from missing days’ supply.

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

**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*)

There are two exclusions for the measure:

* Individuals with cancer
* Individuals with a home INR testing

To examine the effect of these exclusions, the measure rates with and without the exclusions were calculated and compared.

**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*)  
The exclusion was applied to the 10-state data from 2012. The aggregated denominator, numerator, and the measure rate across the 10 states are shown in Table 14.

**Table 14. Measure Rate by Exclusion Status – Home INR Monitoring**

| **Home INR Excluded** | **Denominator** | **Numerator** | **Measure Rate** | **95% CI** |
| --- | --- | --- | --- | --- |
| Yes | 103,025 | 21,345 | 20.7% | 20.5%, 21.0% |
| No | 111,370 | 22,056 | 19.8% | 19.6%, 20.0% |

For the overall denominator, the measure rate excluding patients with home INR is significantly higher than the measure rate including patients with home INR (95% confidence intervals do not overlap). For measure rates including and excluding patients with home INR, the states with the lowest measure rate (Mississippi) and highest measure rates (Rhode Island) are statistically significant different from all other states (p-value ≤0.0353). Texas is statistically different from all other states except Arizona (p-value ≤0.0353) for measure rates including or excluding home INR patients (Table 15).

**Table 15. Exclusion Analysis by States – Home INR Monitoring**

| **State** | **Excluding Patients with Home INR** | | | | **Including Patients with Home INR** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Den** | **Num** | **Rate** | **95% CI** | **Den** | **Num** | **Rate** | **95% CI** |
| All | 103,025 | 21,345 | 20.7% | 20.5%, 21.0% | 111,370 | 22,056 | 19.8% | 19.6%, 20.0% |
| AZ | 4,715 | 939 | 19.9% | 18.8%, 21.1% | 5,336 | 986 | 18.5% | 17.4%, 19.5% |
| DE | 1,424 | 323 | 22.7% | 20.5%, 24.9% | 1,595 | 336 | 21.1% | 19.1%, 23.1% |
| FL | 23,644 | 5,092 | 21.5% | 21.0%, 22.1% | 26,051 | 5,328 | 20.5% | 20.0%, 20.9% |
| IA | 9,719 | 2,267 | 23.3% | 22.5%, 24.2% | 10,020 | 2,295 | 22.9% | 22.1%, 23.7% |
| IN | 12,223 | 2,629 | 21.5% | 20.8%, 22.2% | 13,207 | 2,716 | 20.6% | 19.9%, 21.3% |
| MO | 11,318 | 2,397 | 21.2% | 20.4%, 21.9% | 12,319 | 2,474 | 20.1% | 19.4%, 20.8% |
| MS | 7,572 | 1,218 | 16.1% | 15.3%, 16.9% | 7,986 | 1,251 | 15.7% | 14.9%, 16.5% |
| RI | 1,211 | 388 | 32.0% | 29.4%, 34.7% | 1,284 | 397 | 30.9% | 28.4%, 33.4% |
| TX | 23,173 | 4,217 | 18.2% | 17.7%, 18.7% | 25,128 | 4,357 | 17.3% | 16.9%, 17.8% |
| WA | 8,026 | 1,875 | 23.4% | 22.4%, 24.3% | 8,444 | 1,916 | 22.7% | 21.8%, 23.6% |

For the overall denominator, the measure rate excluding patients with cancer is no different than the measure rate including patients with cancer (95% confidence intervals overlap) (Table 16).

**Table 16. Measure Rate by Exclusion Status – Cancer Diagnosis**

| **Cancer Excluded** | **Denominator** | **Numerator** | **Measure Rate** | **95% CI** |
| --- | --- | --- | --- | --- |
| Yes | 103,025 | 21,345 | 20.7% | 20.5%, 21.0% |
| No | 124,695 | 25,934 | 20.8% | 20.6%, 21.0% |

For measure rates including and excluding patients with a cancer diagnosis, the states with the lowest measure rates (Mississippi) and highest measure rates (Rhode Island) are statistically different from all other states (p-value ≤0.0032). When excluding patients with cancer, Texas is statistically different from all other states (p-value ≤0.0032) except Arizona. However, Texas differs from all states when including patients with cancer (p-value ≤0.0038) (Table 17).

**Table 17. Exclusion Analysis by States – Cancer Diagnosis**

| **State** | **Excluding Patients with Cancer** | | | | **Including Patients with Cancer** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Den** | **Num** | **Rate** | **95% CI** | **Den** | **Num** | **Rate** | **95% CI** |
| All | 103,025 | 21,345 | 20.7% | 20.5%, 21.0% | 124,695 | 25,934 | 20.8% | 20.6%, 21.0% |
| AZ | 4,715 | 939 | 19.9% | 18.8%, 21.1% | 5,872 | 1,202 | 20.5% | 19.4%, 21.5% |
| DE | 1,424 | 323 | 22.7% | 20.5%, 24.9% | 1,761 | 397 | 22.5% | 20.6%, 24.5% |
| FL | 23,644 | 5,092 | 21.5% | 21.0%, 22.1% | 29,948 | 6,490 | 21.7% | 21.2%, 22.1% |
| IA | 9,719 | 2,267 | 23.3% | 22.5%, 24.2% | 11,398 | 2,670 | 23.4% | 22.6%, 24.2% |
| IN | 12,223 | 2,629 | 21.5% | 20.8%, 22.2% | 14,590 | 3,097 | 21.2% | 20.6%, 21.9% |
| MO | 11,318 | 2,397 | 21.2% | 20.4%, 21.9% | 13,535 | 2,858 | 21.1% | 20.4%, 21.8% |
| MS | 7,572 | 1,218 | 16.1% | 15.3%, 16.9% | 8,844 | 1,435 | 16.2% | 15.5%, 17.0% |
| RI | 1,211 | 388 | 32.0% | 29.4%, 34.7% | 1,525 | 502 | 32.9% | 30.6%, 35.3% |
| TX | 23,173 | 4,217 | 18.2% | 17.7%, 18.7% | 27,580 | 5,019 | 18.2% | 17.7%, 18.7% |
| WA | 8,026 | 1,875 | 23.4% | 22.4%, 24.3% | 9,642 | 2,264 | 23.5% | 22.6%, 24.3% |

**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*)  
Statistically significant differences were identified in the measure rate with and without the exclusion of home INR monitoring. Since beneficiaries monitoring INR at home would not have claims for INR tests, this exclusion improves the measures validity. Regarding the exclusion for patients with cancer, it would appear in aggregate that the rates are not significantly different with or without the exclusions. Similarly, at the state level, the confidence intervals overlap between the two versions of the measure. Overall, since cancer patients monitoring rates are not significantly different, we recommend removing this exclusion from the measure, which would reduce the measure complexity.

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

**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?**

X **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**.   
Not applicable

**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*)

Not applicable

**2b4.4. What were the statistical results of the analyses used to select risk factors?**Not applicable

**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*)  
Not applicable

*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*)**:**Not applicable

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
Not applicable

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:  
Not applicable

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

Not applicable

**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*)  
Not applicable

\***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*)

Not applicable

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

**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 identify statistically significant differences in performance, we conducted a comparison of means and percentiles at the state, prescription drug plan, ACO, and physician group levels. Confidence intervals (CI 95%) were calculated around point estimates for each state, prescription drug plan, and ACO and then compared to the overall mean of states, prescription drug plans, and ACOs, respectively. If the confidence intervals did not overlap with the overall mean, the difference was considered statistically significant. (**Note:** Results are not presented for physician groups since reliable comparisons at the physician group level were not feasible).

**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*)  
We analyzed the measure performance by state, Part D plan, and ACO, and the results, along with a discussion of the meaningful differences at each level, are described below.

Meaningful Differences at the State Level – 2008

Two of the eight states (25.0%) had scores statistically significantly lower than the mean, and another four states (50.0%) had scores significantly higher than the mean. Measure rates ranged from 15.8% in Mississippi to 31.9% in Rhode Island, indicating suboptimal performance across all eight states (Table 18).

Meaningful Differences at the State Level – 2012

Two of the 10 states (20.0%) had scores statistically significantly lower than the mean, and another five states (50.0%) had scores significantly higher than the mean. Measure rates ranged from 16.1% in Mississippi to 32.0% in Rhode Island, indicating suboptimal performance across all 10 states (Table 18).

**Table 18. 2008, 2012 State Level Performance**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| 2008 | 8 | 22.5% | 22.0% | 15.8% | 31.9% | 4.6% | 3.2% | 15.8% | 20.5% | 22.0% | 23.7% | 31.9% |
| 2012 | 10 | 22.0% | 21.5% | 16.1% | 32.0% | 4.2% | 3.4% | 17.1% | 19.9% | 21.5% | 23.3% | 27.7% |

Meaningful Differences at the Plan Level

Of the plan scores, 10.0% of providers were statistically significantly lower than the mean, and 20.0% of providers were statistically significantly higher than the mean. For those plans with at least 2,500 eligible episodes, high- (90th percentile) and low- (10th percentile) performing plans were 4.3% apart, indicating suboptimal performance across all plans and variation between high- and low-performing plans (Table 19).

**Table 19. 2012 Part D Plan Level Performance**

| **Year** | **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2012 | 10 | 20.7% | 20.7% | 17.8% | 23.8% | 1.6% | 1.4% | 18.7% | 19.8% | 20.7% | 21.2% | 23.0% |

Meaningful Differences at the ACO Level

Of the ACO scores, 25.0% of providers were statistically significantly lower than the mean, and 16.7% of providers were statistically significantly higher than the mean. For those ACOs with at15 least 180 eligible episodes, high- (90th percentile) and low- (10th percentile) performing ACOs were 15.1% apart, indicating suboptimal performance across all ACOs and variation between high- and low-performing ACOs (Table 20).

**Table 20. 2011 ACO Level Performance**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| 2011 | 24 | 22.6% | 23.0% | 13.0% | 32.9% | 5.5% | 7.9% | 15.2% | 18.3% | 23.0% | 26.2% | 30.4% |

**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 measure results indicate suboptimal performance at all three levels of analysis (ACO, state, and prescription drug plan). The measure is able to clearly distinguish between high- and low-performing providers at the ACO, state, and prescription drug plan levels.

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

**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 datasources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
Not applicable

**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*)  
Not applicable

**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*)  
Not applicable

1. SAS® 9.2 Software; SAS Institute Inc.; Cary, North Carolina [↑](#footnote-ref-2)
2. The 2014 version of the GEM files were used. These files were created by the National Center for Health Statistics (NCHS) and can be found on the Centers for Disease Control and Prevention (CDC) website. [↑](#footnote-ref-3)
3. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM): 2013 release of ICD-10-CM. Available at: [http://www.cdc.gov/nchs/icd/icd10cm.htm. Accessed September 15, 2013](http://www.cdc.gov/nchs/icd/icd10cm.htm.%20Accessed%20September%2015,%202013). [↑](#footnote-ref-4)
4. Microsoft® Excel 2010; Microsoft Corporation; Redmond, Washington [↑](#footnote-ref-5)