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

**Measure Title**: INR Monitoring for Individuals on Warfarin

**Date of Submission**: 10/29/2018

**Type of Measure:** Process

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | ☐ Outcome (*including PRO-PM*) |
| ☐ Cost/resource | X 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, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the 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** demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** **11** demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score. **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; **12**  **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**  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; **14,15** and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** **16** **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| X X administrative claims | X 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:

**PRIOR SUBMISSION**

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

**UPDATED TESTING**

* 2015–2016 administrative claims data from four issuers (referred to as QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4), containing a total of seven Health Insurance Exchange Qualified Health Plan (QHP) products in 2015 and eight in 2016. The following describes the terminology of the units associated with the Health Insurance Exchange: “Issuer” refers to an individual insurance company or insurance organization. The term “product” refers to a package of health coverage benefits that are offered using a particular network type (i.e., health maintenance organization, preferred provider organization, exclusive provider organization, point of service, or indemnity).[1] Unique products for each issuer are referred to using alphabetic labeling (e.g., two unique products from the same issuer are referred to as Product A and Product B).
* 2015–2016 administrative claims data from Medicare Parts A, B, and D for beneficiaries enrolled in stand-alone Part D Prescription Drug Plans (referred to as Medicare PDPs)

Please note that Medicare data were used for measure testing to enhance the measure testing results. At the time this form was completed, CMS does not yet have any plan to add this measure to any quality reporting or value-based purchasing programs for Medicare enrollees but may consider these measures for the future. However, this measure is being considered for use in the Quality Rating System for Qualified Health Plans.

Citation:

1. Centers for Medicare & Medicaid Services. Federal Definitions for Health Insurance Products and Plans. Baltimore, MD: US Department of Health and Human Services; 2016. <https://www.cms.gov/CCIIO/Resources/Training-Resources/Downloads/product-vs-plan-ppt.pdf.> Accessed June 12, 2018.

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

**PRIOR SUBMISSION**

* January 1, 2011 – December 31, 2012

**UPDATED TESTING**

* January 1, 2015 – December 31, 2016

**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 |
| X group/practice | X group/practice |
| ☐ hospital/facility/agency | ☐ hospital/facility/agency |
| X X health plan | X 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*)

**PRIOR SUBMISSION**

Characteristics of the sample for 2011–2012 are summarized in Table 1. All beneficiaries from 10 states (Arizona, Delaware, Florida, Iowa, Indiana, 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 PDP. 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 1. 2011-2012 Sample Characteristics by States, PDPs, 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 2.

**Table 2. 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 |

**UPDATED TESTING**

Characteristics of the data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs are summarized in Tables 3a (2015) and 3b (2016). The data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, and QHP Issuer 4 included all members with claims associated with the QHP products. To align with the 2018 Quality Rating System, Measure Technical Specifications: [1]

* QHP products with 500 or fewer total members were excluded from all analyses, and
* Denominators had to have at least 30 members in order to show the results of analyses.

The 501 member and 30 minimum denominator rules are not part of the measure specifications. The analyses followed these rules to reflect steps that would be taken if the measure were implemented into the Quality Rating System (QHP data).

The Medicare sample included all beneficiaries from the national Medicare claims database who had at least one month of Part A and Part B coverage and no HMO coverage during the year and who were in a stand-alone Medicare PDP. The 501 member and 30 minimum denominator rules were not applied to the Medicare data since the rules are specific to the Quality Rating System (QHP data).

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf.> Accessed July 13, 2018.

**Table 3a. 2015 Sample Characteristics of the Data**

| **Characteristics** | **QHP Issuer 1** | **QHP Issuer 2** | **QHP Issuer 3** | **QHP Issuer 4** | **Medicare PDPs** |
| --- | --- | --- | --- | --- | --- |
| Total Number of QHP Products or Medicare PDPs | 3 | 1 | 2 | 1 | 66 |
| Total Member/Beneficiary Sample Size Enrolled in a QHP Product/PDP | 289,136 | 49,137 | 15,671 | 3,354 | 18,894,628 |
| Mean # of Members/ Beneficiaries per Product/PDP | 96,378 | 49,137 | 7,836 | 3,354 | 286,282 |

**Table 3b. 2016 Sample Characteristics of the Data**

| **Characteristics** | **QHP Issuer 1** | **QHP Issuer 2** | **QHP Issuer 3** | **QHP Issuer 4** | **Medicare PDPs** |
| --- | --- | --- | --- | --- | --- |
| Total Number of QHP Products or Medicare PDPs | 3 | 1 | 3 | 1 | 62 |
| Total Member/Beneficiary Sample Size Enrolled in a QHP Product/PDP | 223,427 | 33,205 | 84,255 | 2,284 | 19,607,672 |
| Mean # of Members/ Beneficiaries per Product/PDP | 74,476 | 33,205 | 28,085 | 2,284 | 316,253 |

**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*)

**PRIOR SUBMISSION**

Demographic characteristics of the beneficiaries in the 2011–2012 datasets are shown in Table 4 below.

**Table 4. 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 5.

**Table 5. 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 (84.34%) |
| **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%) |

**UPDATED TESTING**

Demographic characteristics of members of QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs are shown in Tables 6a (2015) and 6b (2016); however, limited demographic variables were available in our testing data. “N/A” in the tables indicates the data were not available.

**Table 6a. 2015 Demographic Characteristics of Members of QHP Issuers and Medicare PDPs**

| **Characteristics** | **QHP Issuer 1** | **QHP Issuer 2** | **QHP Issuer 3** | **QHP Issuer 4** | **Medicare PDPs** |
| --- | --- | --- | --- | --- | --- |
| **Total Sample Size** | 289,136 | 49,137 | 15,671 | 3,354 | 18,894,628 |
| **Sex n (% of Total Sample)\*** | | | | | |
| Female | 150,116  (51.9) | 21,399  (43.5) | 7,043  (44.9) | 1,538  (45.9) | 10,413,926  (55.1) |
| Male | 139,020  (48.1) | 27,738  (56.5) | 8,628  (55.1) | 1,816  (54.1) | 8,480,702  (44.9) |
| **Age n (% of Total Sample)\*** | | | | | |
| <18 years | 9,584  (3.3) | 3,600  (7.3) | 1,578  (10.1) | 247  (7.4) | 132  (0.0) |
| 18–26 years | 38,590  (13.4) | 3,633  (7.4) | 1,640  (10.5) | 333  (9.9) | 105,869  (0.6) |
| 27–44 years | 81,098  (28.0) | 12,486  (25.4) | 5,671  (36.2) | 1,022  (30.5) | 911,610  (4.8) |
| 45–64 years | 152,252  (52.7) | 28,965  (59.0) | 6,603  (42.1) | 1,711  (51.0) | 2,958,692  (15.7) |
| ≥65 years | 7,612  (2.6) | 453  (0.9) | 179  (1.1) | 41  (1.2) | 14,918,325  (79.0) |
| **Race n (% of Total Sample)\*** | | | | | |
| White/ Caucasian | N/A | N/A | N/A | N/A | 15,782,130  (83.5) |
| African-American | N/A | N/A | N/A | N/A | 1,893,242  (10.0) |
| Hispanic | N/A | N/A | N/A | N/A | 383,461  (2.0) |
| Other | N/A | N/A | N/A | N/A | 633,329  (3.4) |
| Unknown | N/A | N/A | N/A | N/A | 202,466  (1.1) |

**\***Numbers in parentheses represent the column percent by demographic characteristic.

**Table 6b. 2016 Demographic Characteristics of Members of QHP Issuers and Medicare PDPs**

| **Characteristics** | **QHP Issuer 1** | **QHP Issuer 2** | **QHP Issuer 3** | **QHP Issuer 4** | **Medicare PDPs** |
| --- | --- | --- | --- | --- | --- |
| **Total Sample Size** | 223,427 | 33,205 | 84,255 | 2,284 | 19,607,672 |
| **Sex n (% of Total Sample)\*** | | | | | |
| Female | 116,111  (52.0) | 14,546  (43.8) | 38,433  (45.6) | 1,027  (45.0) | 10,787,561  (55.0) |
| Male | 107,316  (48.0) | 18,659  (56.2) | 45,822  (54.4) | 1,257  (55.0) | 8,820,111  (45.0) |
| **Age n (% of Total Sample)\*** | | | | | |
| <18 years | 8,536  (3.8) | 3,077  (9.3) | 8,618  (10.2) | 207  (9.1) | 121  (0.0) |
| 18–26 years | 27,732  (12.4) | 2,445  (7.4) | 8,268  (9.8) | 236  (10.3) | 101,020  (0.5) |
| 27–44 years | 58,419  (26.2) | 8,584  (25.8) | 27,730  (32.9) | 724  (31.7) | 888,545  (4.5) |
| 45–64 years | 121,304  (54.3) | 18,756  (56.5) | 38,748  (46.0) | 1,089  (47.7) | 2,942,822  (15.0) |
| ≥65 years | 7,436  (3.3) | 343  (1.0) | 891  (1.1) | 28  (1.2) | 15,675,164  (79.9) |
| **Race n (% of Total Sample)\*** | | | | | |
| White/ Caucasian | N/A | N/A | N/A | N/A | 16,355,081 (83.4) |
| African-American | N/A | N/A | N/A | N/A | 1,920,626 (9.8) |
| Hispanic | N/A | N/A | N/A | N/A | 402,787 (2.1) |
| Other | N/A | N/A | N/A | N/A | 673,534 (3.4) |
| Unknown | N/A | N/A | N/A | N/A | 255,644 (1.3) |

**\***Numbers in parentheses represent the column percent by demographic characteristic.

**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**.

**UPDATED TESTING**

The following sources of data were used in testing NQF 0555 at the health plan level:

1. QHP products: claims data from issuers, consisting of hospital and office visit, pharmacy, and laboratory claims (when available); enrollment data; members’ demographic data; and provider information.
2. Medicare: claims data from Medicare Parts A and B and stand-alone Part D PDPs, consisting of inpatient and outpatient claims and prescription drug events; enrollment data; members’ demographic data; and provider information.

The difference in the data used for the various aspects of testing is shown in Table 7. “X” indicates no data were available.

**Table 7. Data Used to Test the Measure**

| **Testing of the Measure** | **QHP Data** | **Medicare Data** |
| --- | --- | --- |
| Development of the Denominator |  |  |
| Development of the Numerator |  |  |
| Data Element Feasibility |  |  |
| Measure Performance Reliability (Signal to Noise) |  |  |
| Calculating Measure Performance |  |  |
| Convergent Validity | X |  |
| Exclusion Analyses |  |  |
| Disparities Analyses |  |  |

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

**UPDATED TESTING**

This process measure, NQF 0555, is not risk adjusted and therefore an analysis of social risk factors was not conducted.

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

**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*)

**PRIOR SUBMISSION**

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], 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 on 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.

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.

**UPDATED TESTING**

Measure score reliability was estimated using a beta-binomial model. For the QHP data, the mean reliability was calculated across QHP products. Reliability estimates for Medicare PDPs were computed by using the methods of minimum denominator and volume categories, described by Scholle et al. (2008).[1] This difference in approach to the data is due to the limited number of available QHP products.

Reliability, QHP Products, Issuer 1, Issuer 2, and Issuer 3

We calculated reliability for each QHP product and the mean reliability across QHP products in 2016. Note that QHP Issuer 4 did not have sufficient denominator sizes for analyses and is thus not presented in the results section for reliability, below. Sufficient denominator size for display was defined as 30 members or more in the denominator to align with the 2018 Quality Rating System Measure Technical Specifications.[2]

Minimum Denominator for Reliability, Medicare PDPs

The testing conducted for this comprehensive re-evaluation used the same methods for the 2016 Medicare PDP sample as described above with the exception that we used the method of minimum denominator and volume categories from Scholle et al. instead of the mean denominator.[2] This method assumes that the denominator size in each volume category is equal to the minimum for that category. As such, it provides a more conservative estimate of reliability for each volume category.

Citations:

1. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. *Am J Manag Care.* 2008;14(12):833-838.

2. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf.> Accessed July 13, 2018.

**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*)

**PRIOR SUBMISSION**

We conducted reliability tests across measurement units, and the results from the state level, including reliability statistics and assessments of adequacy, are provided below.

We concluded that the reliability test was adequate, since all state-level reliability scores were greater than 0.7, indicating that the measure would produce reliable scores at the state level (Table 8).

**Table 8. 2011-2012 State Reliability and Assessment of Adequacy for Tests Conducted**

| **State** | **Measure Rate (Reliability)** |
| --- | --- |
| AZ | 74.62% (0.99) |
| DE | 75.45% (0.99) |
| FL | 74.28% (0.99) |
| IA | 83.19% (0.99) |
| IN | 77.81% (0.99) |
| MO | 76.30% (0.99) |
| MS | 65.53% (0.99) |
| RI | 88.61% (0.99) |
| TX | 64.28% (0.99) |
| WA | 78.39% (0.99) |

Using the method of mean denominator and volume categories, a minimum denominator of 100 individuals resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between plans. Furthermore, more than half (52.0%) of the plans with at least one patient attributed (n=75) had at least 100 individuals in the measure denominator and a reliable score (Table 9).

**Table 9. 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)** | **Mean Rate of Plans** | **Reliability Score** |
| 2012 | 100 | 39 (52.0%) | 74.52% | 0.71 |

**UPDATED TESTING**

Reliability, QHP Products, Issuer 1, Issuer 2, and Issuer 3

Among the QHP products tested, reliability ranged from 0.60 to 0.79 with a mean reliability of 0.70 (Table 10a), which suggests sufficient signal relative to noise to discriminate performance between plans.

**Table 10a. 2016 Reliability Among QHP Products with At Least 30 Members in the Denominator**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **QHP Issuer** | **Product** | **Denominator** | **Numerator** | **Measure Rate** | **Variance Within** | **Variance Between** | **Reliability Score** |
| Issuer 1 | B | 326 | 143 | 43.9% | 7.55 | 29.19 | 0.79 |
| Issuer 2 | A | 203 | 120 | 59.1% | 11.91 | 29.19 | 0.71 |
| Issuer 3 | A | 185 | 105 | 56.8% | 13.27 | 29.19 | 0.69 |
| Issuer 3 | B | 126 | 71 | 56.4% | 19.52 | 29.19 | 0.60 |
| **Mean** |  |  |  |  |  |  | **0.70** |

Minimum Denominator for Reliability, Medicare PDPs

Using the method of minimum denominator and volume categories, a minimum of 100 members in the denominator results in an overall reliability score of 0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between units of analysis.[1] Of the 61 PDPs in 2016, the majority (83.6%) of PDPs had at least 100 individuals in the measure denominator, representing a mean performance rate of 71.74% (reliability = 0.70) (Table 10b).

**Table 10b. 2016 Medicare PDP Reliability and Assessment of Adequacy for Tests Conducted**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Min Denominator** | **Total # of PDPs** | **# of PDPs with at Least 100 Individuals** | **Mean Rate of Plans with at Least 100 Individuals** | **Reliability Score** |
| 100 | 61 | 51 | 71.74% | 0.70 |

Citation:

1. Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling—reliability and risk of misclassification. *N Engl J Med.* 2010;362(11):1014-1021. doi: 10.1056/NEJMsa0906323.

**PRIOR SUBMISSION**

Using the method of mean denominator and volume categories, a minimum denominator of 50 individuals measured resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between physician groups. Approximately 11% of physician groups with at least one patient attributed (n=6,594) had at least 50 individuals in the measure denominator and a reliable score (Table 11).

**Table 11. 2012 Physician Group Reliability and Assessment of Adequacy for Tests Conducted**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Min Denominator** | **# of Physician Groups (% of physician groups with at least 1 individual attributed)** | **Mean Rate of Physician Groups** | **Reliability Score** |
| 2012 | 50 | 739 (11.21%) | 75.66% | 0.73 |

Using the method of mean denominator and volume categories, a minimum denominator of 50 individuals resulted in an overall reliability score of >0.7, which is within acceptable norms and indicates sufficient signal strength to discriminate performance between ACOs (Table 12). The aforementioned criteria resulted in 100.0% of all ACOs (31 of 31 ACOs) with reliable scores (Table 13).

**Table 12. 2011 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 | 50 | 31 (100.0%) | 75.34% | 0.71 |

**Table 13. 2011 Individual ACO Reliability and Assessment of Adequacy for Tests Conducted**

| **ACO #** | **Denominator** | **Measure Rate (Reliability)** |
| --- | --- | --- |
| 1 | 1,124 | 85.14% (0.99) |
| 2 | 650 | 87.08% (0.98) |
| 3 | 1,034 | 80.37% (0.98) |
| 4 | 443 | 70.65% (0.95) |
| 5 | 466 | 64.38% (0.95) |
| 6 | 657 | 81.13% (0.97) |
| 7 | 502 | 65.54% (0.95) |
| 8 | 947 | 63.99% (0.97) |
| 9 | 848 | 58.49% (0.97) |
| 10 | 394 | 69.80% (0.94) |
| 11 | 1,451 | 84.01% (0.99) |
| 12 | 402 | 78.61% (0.96) |
| 13 | 1,022 | 85.62% (0.97) |
| 14 | 902 | 81.71% (0.98) |
| 15 | 943 | 83.67% (0.98) |
| 16 | 1,609 | 88.56% (0.98) |
| 17 | 1,175 | 81.28% (0.99) |
| 18 | 473 | 79.28% (0.99) |
| 19 | 450 | 78.00% (0.96) |
| 20 | 815 | 90.67% (0.96) |
| 21 | 697 | 89.53% (0.99) |
| 22 | 137 | 69.34% (0.98) |
| 23 | 1,577 | 74.51% (0.97) |
| 24 | 825 | 69.33% (0.95) |
| 25 | 440 | 62.95% (0.93) |
| 26 | 350 | 65.71% (0.96) |
| 27 | 361 | 62.60% (0.98) |
| 28 | 448 | 76.34% (0.95) |
| 29 | 481 | 61.12% (0.95) |
| 30 | 772 | 84.59% (0.98) |
| 31 | 2,003 | 61.66% (0.99) |

**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?*)

**PRIOR SUBMISSION**

The results indicated that the measure, as currently specified, was reliable at the state, prescription drug plan, and ACO levels. However, due to sample size issues only a small percentage of physician groups (11.21%) have an adequate number of patients for reliable measurement.

**UPDATED TESTING**

Reliability, QHP Products, Issuer 1, Issuer 2, and Issuer 3

The results indicate that NQF 0555 is reliable at the health plan level, based on a sample of QHP products. Among the products with at least 30 denominator members, the average reliability was 0.70, which suggests sufficient signal relative to noise to discriminate performance between plans.

Reliability, Medicare PDPs

The results indicate that NQF 0555 is reliable at the health plan level, based on Medicare PDP data with at least 100 members in the denominator. In 2016, the majority of Medicare PDPs (83.6%) had at least 100 members in the denominator, which produced measure performance rates with sufficient reliability (0.70) to distinguish differences in performance among plans.

Based on the larger sample from the Medicare data, the reliability findings suggest that a denominator size of at least 100 members would be needed to achieve reliable results at the health plan level.

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

**2b1. VALIDITY TESTING**

**2b1.1. What level of validity testing was conducted**? (*may be one or both levels*)  
☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

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*)

**2b1.2. 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)*  
**PRIOR SUBMISSION**

Performance Measure Score

1. Convergent Validity - Relationship to another measure as expected (NQF 0555 compared to NQF 0556), Pearson Correlation Score
2. Systematic Assessment of Face Validity, Likert Scale, Overall Mean and Median Score (Discussed in 2.b2.3)
3. Threats to Validity, Analysis of Missing Data, Frequency

Convergent Validity: We compared a related NQF-endorsed measure, NQF 0556, which assesses INR monitoring after an interacting anti-infective drug is prescribed. 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.

**UPDATED TESTING**

Convergent Validity: Using Pearson’s correlation coefficients, we compared the performance of NQF 0555 with NQF 0541 (Proportion of Days Covered [PDC]: 3 Rates by Therapeutic Category), which has three rates of medication adherence and is part of the Medicare Part D Star Rating Program in 2015. NQF 0541 assesses adherence to medications for diabetes, hypertension, and cholesterol reduction (*Medication Adherence for Diabetes Medications, Medication Adherence for Hypertension [RAS Antagonists],* and *Medication Adherence for Cholesterol [Statins]*). Our rationale for this comparison is as follows: plans with higher performance on medication adherence should have similar performance with INR testing, since both measures assess appropriate medication management.

**PRIOR SUBMISSION**

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 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).

**UPDATED TESTING**

Face Validity: We systematically evaluated the face validity of NQF 0555 and the measure score after testing was completed. The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®) with the Technical Expert Panel (TEP) advising the project. The TEP is composed of three representatives from large QHP issuers and nine representatives from other stakeholder groups, such as measurement industry representatives, clinical and nonclinical experts, and patient/caregiver representatives. TEP members were specifically asked whether they agree with the following statement: “The performance scores resulting from the measure NQF 0555 INR Monitoring for Individuals on Warfarin, as specified, can be used to distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy.” They responded “yes” or “no,” indicating either they did or they did not agree with the previous statement.

**PRIOR SUBMISSION**

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.

**UPDATED TESTING**

Threats to Validity: We examined the missingness of the prescription variable, *days’ supply*, in our data.

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

**PRIOR SUBMISSION**

Convergent Validity: The measure rate is positively correlated with the NQF-endorsed measure, INR for Individuals Taking Warfarin and Interacting Anti-Infective Medications (NQF 0556) at the ACO level (ρ=0.745, p<0.0001). The distribution of the measure rates is presented in Table 14.

**Table 14. 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% |

**UPDATED TESTING**

Convergent Validity: Results for NQF 0541 were available for 57 PDPs for the diabetes adherence rate and 58 PDPs for the hypertension and cholesterol medication adherence rates. The analysis revealed significant relationships between NQF 0555 measure scores and all three rates of medication adherence (p<0.0001 for all correlations; diabetes: r=0.591, hypertension: r=0.700, cholesterol: r=0.751). These results indicate positive linear associations with large effect sizes between NQF 0555 and three independent measure rates of medication adherence at the PDP level of analysis (Figures 1-3). According to Cohen’s thresholds for product-moment correlations, 0.50 or higher is considered a large correlation.[1]

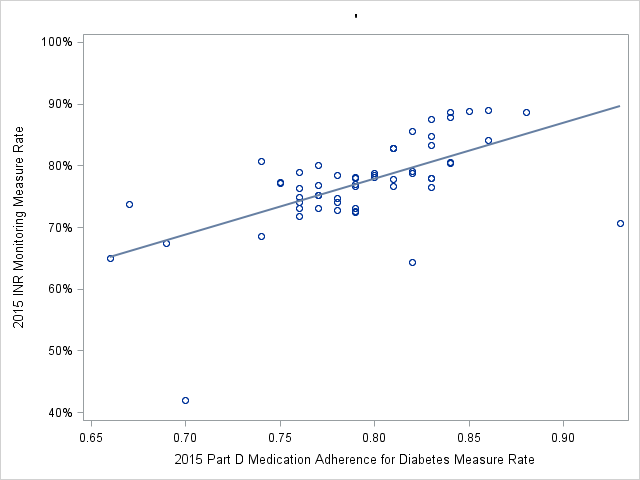
Citation:

1. Cohen J. A power primer. *Psychol Bull.* 1992;112(1):155-159.

**Figure 1. Association Between Performance Rates for NQF 0555**

**and *Medication Adherence for Diabetes Medications*, Medicare**

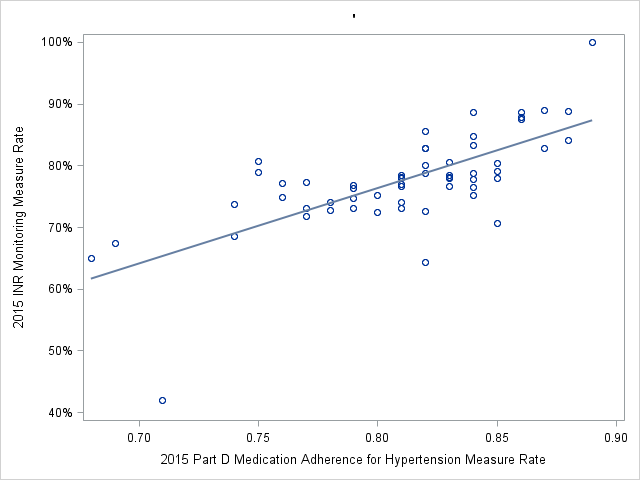
**PDPs, 2015**



**Figure 2. Association Between Performance Rates for NQF 0555**

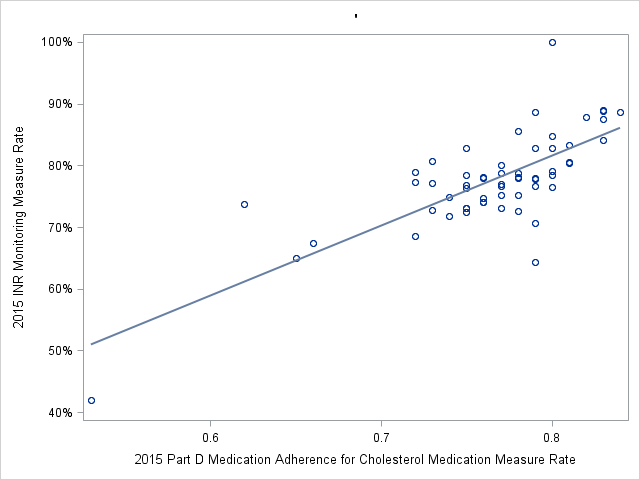
**and *Medication Adherence for Hypertension (RAS Antagonists)*,**

**Medicare PDPs, 2015**



**Figure 3. Association Between Performance Rates for NQF 0555 and**

***Medication Adherence for Cholesterol (Statins)*, Medicare PDPs, 2015**



**PRIOR SUBMISSION**

Systematic Assessment of Face Validity: Fifteen of the 21 (71.4 %) 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 15.

**Table 15. Results of the Face Validity Evaluation**

|  |  |
| --- | --- |
| **Rating** | **Number of TEP (%)** |
| 5 (Strongly Agree) | 4 (26.7%) |
| 4 (Agree) | 8 (53.3%) |
| 3 (Neutral) | 2 (13.3%) |
| 2 (Disagree) | 1 (6.7%) |
| 1 (Strongly Disagree) | 0 |

Of the TEP members who evaluated the measure for face validity, 80% (12/15) strongly agreed or agreed that the measure was valid as specified. The mean rate was 4, and the median rate was 4.

**UPDATED TESTING**

Systematic Assessment of Face Validity: Nine out of nine TEP members (100%) responding to the face validity survey agreed that NQF 0555 was valid as specified. Three TEP members did not complete the survey.

**PRIOR SUBMISSION**

Threats to Validity: Percentage of individuals in the denominator with one or more claims with missing days’ supply - 0/263,080 (0%).

**UPDATED TESTING**

Threats to Validity: No individuals in either the QHP or Medicare PDP denominators had missing days’ supply.

**2b1.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

**PRIOR SUBMISSION**

Convergent Validity: The measure rates between NQF 0555 and NQF 0556 were strongly correlated (>0.7) as expected, and this adds further support that the measure as specified is valid.

**UPDATED TESTING**

Convergent Validity: Performance comparison between NQF 0541, representing three rates of medication adherence, and NQF 0555 was strongly and positively correlated at the PDP level. The results support our hypothesized relationship between NQF 0555 and NQF 0541 and demonstrate that NQF 0555 is valid in capturing the quality of care related to medication management.

**PRIOR SUBMISSION**

Face Validity: In summary, 80% of TEP members who responded to the survey strongly agreed or agreed that the measure has face validity.

**UPDATED TESTING**

Face Validity: Of the TEP members who responded to the survey, 100% agreed that NQF 0555 can be used to distinguish good from poor plan-level quality related to the process of administering at least one INR monitoring test during each 56-day interval among those with active warfarin therapy.

**PRIOR SUBMISSION**

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.

**UPDATED TESTING**

Threats to Validity: Our evaluation of the days’ supply field in both the QHP and Medicare PDP data resulted in zero missing values. Therefore, we conclude that missing data are not a threat to validity.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA** ☐ **no exclusions — *skip to section 2b4***

**2b2.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

**PRIOR SUBMISSION**

Individuals with a home INR testing are excluded. To examine the effect of this exclusion, the measure rates with and without the exclusion were calculated and compared.

**UPDATED TESTING**

Individuals with home INR monitoring are excluded from the NQF 0555 denominator because not all of their INR tests are reliably captured in claims. The INR tests conducted at home are not submitted as individual claims. Therefore, the frequency of the INR tests cannot be ascertained for this population, which prohibits determining whether a home INR test was conducted within the 56-day timeframe specified by the numerator of this measure. To examine the effect of this exclusion, the measure rates with and without the exclusion were calculated and compared using data from QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, QHP Issuer 4, and Medicare PDPs. QHP Issuer 4 did not have sufficient denominator sizes (n=30) for analyses and is thus not included in the results shown in Table 18.[1]

Citation:

1. Centers for Medicare & Medicaid Services. 2018 Quality Rating System Measure Technical Specifications. Baltimore. MD: US Department of Health and Human Services; 2018. <https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/Downloads/Revised_QRS-2018-Measure-Tech-Specs_20170929_508.pdf.> Accessed July 13, 2018.

2015-2017 INR MONITORING AT HOME HCPCS CODES FOR EXCLUSION:

G0248 – Demonstrate Use Home INR Mon

G0249 – Provide Test Mats & Equip Home INR

G0250 – MD INR Test Review Inter Mgmt

**2b2.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)  
**PRIOR SUBMISSION**

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 below in Table 16. In addition, Table 17 shows the results by states.

**Table 16. Measure Rate by Exclusion Status**

| **Home INR Excluded** | **Denominator** | **Numerator** | **Measure Rate** | **95% CI** |
| --- | --- | --- | --- | --- |
| **Yes** | 263,080 | 193,606 | 73.6% | 73.4%, 73.8% |
| **No** | 281,812 | 196,757 | 69.8% | 69.6%, 70.0% |

**Table 17. Exclusion Analysis by States**

| **State** | **Excluding Patients with Home INR** | | | | **Including Patients with Home INR** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Den** | **Num** | **Rate** | **95% CI** | **Den** | **Num** | **Rate** | **95% CI** |
| All | 263,080 | 193,606 | 73.6% | 73.4%, 73.8% | 281,812 | 196,757 | 69.8% | 69.6%, 70.0% |
| AZ | 13,217 | 9,863 | 74.6% | 73.9%, 75.4% | 14,731 | 10,123 | 68.7% | 68.0%, 69.5% |
| DE | 4,028 | 3,039 | 75.5% | 74.1%, 76.8% | 4,371 | 3,091 | 70.7% | 69.4%, 72.1% |
| FL | 64,685 | 48,048 | 74.3% | 73.9%, 74.6% | 70,384 | 49,081 | 69.7% | 69.4%, 70.1% |
| IA | 23,399 | 19,466 | 83.2% | 82.7%, 83.7% | 23,979 | 19,554 | 81.6% | 81.1%, 82.0% |
| IN | 30,056 | 23,388 | 77.8% | 77.3%, 78.3% | 32,261 | 23,714 | 73.5% | 73.0%, 74.0% |
| MO | 27,245 | 20,787 | 76.3% | 75.8%, 76.8% | 29,290 | 21,093 | 72.0% | 71.5%, 72.5% |
| MS | 17,513 | 11,476 | 65.5% | 64.8%, 66.2% | 18,373 | 11,654 | 63.4% | 62.7%, 64.1% |
| RI | 3,828 | 3,392 | 88.6% | 87.6%, 89.6% | 4,051 | 3,445 | 85.0% | 83.9%, 86.1% |
| TX | 55,761 | 35,845 | 64.3% | 63.9%, 64.7% | 60,031 | 36,501 | 60.8% | 60.4%, 61.2% |
| WA | 23,348 | 18,302 | 78.4% | 77.9%, 78.9% | 24,341 | 18,501 | 76.0% | 75.5%, 76.5% |

For the overall cohort, 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 home INR, there is a statistically significant difference between all pairwise comparisons of states (p ≤ 0.05) except for Arizona, Delaware, Florida, and Missouri.

**UPDATED TESTING**

To determine the effect of the exclusion on the 2016 NQF 0555 measure rates, the rates were calculated with and without the exclusion, as shown in Table 18.

**Table 18. 2016 INR Measure Rate by Exclusion Status**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Product** | **Exclusion Status** | **Denominator** | **Numerator** | **Measure Rate** | **95% CI** |
| **QHP Issuer 1** | | | | | |
| B | **No Exclusions** | 328 | 143 | 43.6% | 38.1%, 49.1% |
| B | **Home INR Monitoring Excluded** | 326 | 143 | 43.9% | 38.3%, 49.4% |
| **QHP Issuer 2** | | | | | |
| A | **No Exclusions** | 205 | 122 | 59.5% | 52.8%, 66.2% |
| A | **Home INR Monitoring Excluded** | 203 | 120 | 59.1% | 52.4%, 65.9% |
| **QHP Issuer 3** | | | | | |
| A | **No Exclusions** | 185 | 105 | 56.8% | 49.6%, 63.9% |
| A | **Home INR Monitoring Excluded** | 185 | 105 | 56.8% | 49.6%, 63.9% |
| B | **No Exclusions** | 126 | 71 | 56.4% | 47.7%, 65.0% |
| B | **Home INR Monitoring Excluded** | 126 | 71 | 56.4% | 47.7%, 65.0% |
| **Medicare PDPs** | | | | | |
|  | **No Exclusions** | 1,187,597 | 771,073 | 64.9% | 64.8%, 65.0% |
|  | **Home INR Monitoring Excluded** | 1,059,826 | 754,993 | 71.2% | 71.2%, 71.3% |

**2b2.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis. Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)  
**PRIOR SUBMISSION**

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.

**UPDATED TESTING**

Individuals with home INR monitoring are excluded from the NQF 0555 denominator because not all of their INR tests are reliably captured in the claims. The INR tests conducted at home are not submitted as individual claims. Furthermore, although two of the HCPCS codes used to identify home monitoring are for provision of INR test materials and physician review of test results, these two codes can be associated with up to four INR tests per claim. Therefore, the frequency of the INR tests cannot be accurately ascertained for this population. Our empirical analysis confirm that measure rates were lower in the Medicare population if the exclusion for beneficiaries was not applied because beneficiaries could meet the denominator definition of being on warfarin therapy for at least 56 days but did not meet the numerator since their monitoring of INR was conducted at home. For the QHP data, the rates did not differ significantly if the exclusion for members was not applied, because only a small number of individuals were conducting home INR monitoring. Therefore, we have retained the measure exclusion, since patients form either population monitoring INR at home would not have reliable claims data for INR tests that could be used to satisfy the measure specifications (a test every 56 days).

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

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.***

**2b3.1. 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

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

**2b3.2. 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

**2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any “ordering” of risk factor inclusion**; for example, are social risk factors added after all clinical factors? Not applicable

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

**Published literature**

**Internal data analysis**

**Other (please describe)**

**2b3.4a. What were the statistical results of the analyses used to select risk factors?** Not applicable

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

**2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)  
*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***if stratified, skip to 2b4.9*** Not applicable

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:** Not applicable

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*): Not applicable

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**: Not applicable

**2b3.9. Results of Risk Stratification Analysis**: Not applicable

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*) Not applicable

**2b3.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*  
**PRIOR SUBMISSION**

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

**UPDATED TESTING**

For this comprehensive re-evaluation, we used the same methods as described above for the evaluation of performance variation among QHP products and Medicare PDPs.

**2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
**PRIOR SUBMISSION**

Meaningful Differences at the State Level – 2012: Two of the 10 states (20.0%) had scores statistically significantly lower than the mean, and the other eight states (80.0%) had scores significantly higher than the mean. Measure rates ranged from 64.3% in Texas to 88.6% in Rhode Island, indicating suboptimal performance across all eight states (Table 19).

**Table 19. 2012 State Level Performance**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| 10 | 75.8% | 75.9% | 64.3% | 88.6% | 7.2% | 4.1% | 64.9% | 74.3% | 75.9% | 78.4% | 85.9% |

Meaningful Differences at the Plan Level – 2012: Of the plan scores, 33.3% of providers were statistically significantly lower than the mean, and 51.3% of providers were statistically significantly higher than the mean. For those plans with at least 100 eligible individuals, high- (90th percentile) and low- (10th percentile) performing plans were 18.9% apart, indicating suboptimal performance across plans and variations between high- and low-performing plans (Table 20).

**Table 20. 2012 Prescription Drug Plan Level Performance**

| **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 39 | 74.5% | 75.6% | 59.7% | 88.3% | 7.2% | 12.6% | 64.8% | 68.5% | 75.6% | 81.0% | 83.6% |

**UPDATED TESTING**

Meaningful Differences at the Plan Level – 2016

Measure rates across QHP products ranged from 43.9% to 59.1% (Table 21a) with a mean measure rate of 54.0%. These rates were substantially lower than the rate observed among Medicare PDPs (71.7%) (Table 21b).

**Table 21a. 2016 QHP Performance for Those with at Least 30 Members in the Denominator**

| **QHP Issuer** | **QHP Product** | **Rate** | **Confidence Interval** |
| --- | --- | --- | --- |
| 1 | B | 43.9% | 38.3%, 49.4% |
| 2 | A | 59.1% | 52.4%, 65.9% |
| 3 | A | 56.8% | 49.6%, 63.9% |
| 3 | B | 56.4% | 47.7%, 65.0% |

The reliability findings suggested that a denominator size of at least 100 members would be needed to achieve reliable results at the health plan level. Therefore, among Medicare PDPs with at least 100 denominator beneficiaries, we found that 41.2% (21/51) of plans had rates significantly lower than the mean, and 37.3% (19/51) of plans had rates significantly greater than the mean. For PDPs with at least 100 members, the difference in performance between high-performing (i.e., 90th percentile) and low-performing (i.e., 10th percentile) PDPs was 18.2%, indicating both variation between high- and low-performing PDPs and suboptimal performance across PDPs (Table 21b).

**Table 21b. 2016 Medicare PDP Performance for Those with at Least 100 Members in the Denominator**

| **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 51 | 71.7% | 71.4% | 46.4% | 85.1% | 7.5% | 10.1% | 64.0% | 67.3% | 71.4% | 77.4% | 82.2% |

**PRIOR SUBMISSION**

Meaningful Differences at the Physician Group Level – 2012: Of the physician group scores, 24.4% of providers were statistically significantly lower than the mean, and 28.1% of providers were statistically significantly higher than the mean, indicating a wide range of scores. For those physician groups with at least 50 eligible individuals, high- (90th percentile) and low- (10th percentile) performing physician groups were 28.4% apart. The results indicate ample room for improvement and meaningful differences in quality of care between the highest and lowest performing physician groups (Table 22).

**Table 22. 2012 Physician Group Level Performance**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| 739 | 75.7% | 77.2% | 14.8% | 97.1% | 11.2% | 15.4% | 60.3% | 68.9% | 77.2% | 84.2% | 88.7% |

Meaningful Differences at the ACO Level: Of the ACO scores, 41.9% of providers were statistically significantly lower than the mean, and 41.9% of providers were statistically significantly higher than the mean. For those ACOs with at least 50 eligible individuals, high- (90th percentile) and low- (10th percentile) performing ACO were 24.5% apart, indicating suboptimal performance across ACOs and variation between high- and low-performing ACOs (Table 23).

**Table 23. ACO Level Performance (2011)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **Mean** | **Median** | **Min** | **Max** | **STD** | **IQR** | **P10** | **P25** | **P50** | **P75** | **P90** |
| 31 | 75.3% | 78.0% | 58.5% | 90.7% | 9.8% | 18.5% | 62.6% | 65.5% | 78.0% | 84.0% | 87.1% |

**2b4.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)  
**PRIOR SUBMISSION**

The overall mean of ~75% of patients having an INR test every 56 days indicates that measure performance is suboptimal. Furthermore, across measurement units, there was ample variation in performance between high- and low-performing plans indicating room for improvement in INR monitoring rates.

**UPDATED TESTING**

The low performance rates of the QHP products (average rate of 54.0% in 2016) suggests substantial opportunity for improvement in the management of patients on warfarin among QHPs in the Health Insurance Exchanges. Among Medicare PDPs, measure rates decreased from 2012 to 2016. In 2016, there was variation among Medicare PDP measure rates, and measure performance remained suboptimal (average rate of 71.7%) among Medicare PDPs. The performance rates of this measure suggest opportunity for improving care for QHP consumers and Medicare beneficiaries on warfarin therapy.

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

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

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

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

**2b5.1. Describe the method of testing conducted to 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

**2b5.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
Not applicable

**2b5.3. 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

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

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

**2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)

**UPDATED TESTING**

NQF 0555 is a claims-based measure and relies on final paid claims from payors (Medicare, QHP Issuer 1, QHP Issuer 2, QHP Issuer 3, or QHP Issuer 4). The most critical data element that could lead to missing cases, *days’ supply of medication*, was complete in the datasets used for testing. None of the claims contained missing data for the element *days’ supply of medication*.

**2b6.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)  
Not applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Not applicable