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

**Measure Number** (*if previously endorsed*)**:** 0022

**Measure Title**: Use of High-Risk Medications in Older Adults

**Date of Submission**: 11/2/2020

**Type of Measure:**

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

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b1, 2b2, and 2b4 must be completed.** * **For outcome and resource use measures**, section **2b3** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b5** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b1-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for version 7.1 of the Measure Testing Attachment. |

|  |
| --- |
| **Note:** The information provided in this form is intended to aid the Standing Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **instrument-based measures** (including PRO-PMs) **and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b1.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **instrument-based measures (including PRO-PMs) and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b2.** **Exclusions** are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b3.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b4.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b5.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b6.** Analyses identify the extent and distribution of **missing data** (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality. The degree of consensus and any areas of disagreement must be provided/discussed.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.17*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| claims | claims |
| registry | registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: Click here to describe | other: Click here to describe |

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

**2020 Submission**

N/A

**2016 Submission**

N/A

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

**2020 Submission**

HEDIS Health Plan performance data from 2018

**2016 Submission**

HEDIS Health Plan performance data from 2012-2014

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

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.20*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Click here to describe | other: Click here to describe |

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

**2020 Submission**

This measure assesses whether Medicare members ages 65 years and older had at least 2 dispensing events for the same high-risk medication. Testing was completed at the health plan level which is appropriate for the level of reporting for this measure.

Measure score reliability testing and construct validity testing: Data used to assess reliability and validity were calculated from all Medicare plans submitting data to NCQA in 2018 for this HEDIS measure. This data came from 502 Medicare plans in total that were geographically diverse and varied in size.

Systematic evaluation of face validity:

The measure was assessed for face validity with three independent panels of experts.

* The Geriatric Measurement Advisory Panel includes 15 experts in geriatric health, including representation by consumers, health plans, health care providers, and policy makers.
* The Technical Measurement Advisory Panel includes 12 members, including representation by health plan methodologists, clinicians, HEDIS auditors and state/federal users of measures.
* NCQA’s Committee on Performance Measurement (CPM) oversees measures used in NCQA programs and includes representation by purchasers, consumers, health plans, health care providers, and policy makers. This panel is composed of 17 independent members that reflect the diversity of constituencies that performance measurement serves. The CPM’s recommendations are reviewed and approved by NCQA’s Board of Directors.

**2016 Submission**

Validity statistics were calculated from 2014 HEDIS Health Plan performance data that included 488 Medicare health plans. This included all Medicare health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size. The average (mean) eligible population for this measure across health plans was 22,043.

**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*)   
**2020 Submission**

HEDIS data are summarized at the health plan level for the Medicare product line. Below is a description of the sample. It includes number of health plans submitting measure data for HEDIS, as well as the average and median eligible population for the measure across health plans.

Table 1. Mean and median eligible population for *Use of High-Risk Medications in Older Adults,* 2018

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Number of Plans** | **Mean number of eligible members per plan** | **Median number of eligible members per plan** |
| 2018 | 502 | 28,463 | 5,893 |

**2016 Submission**

*This question was not on the 2016 form.*

**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**.

**2020 Submission**

No differences in the data used for reliability and construct validity testing. The systematic assessment of face validity was done with multiple multi-stakeholder expert panels as described in Section 1.5 above.

**2016 Submission**

*This question was not on the 2016 form.*

**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.

**2020 Submission**

We did not assess data by social risk factors. Social risk factor data were not available in reported results. This measure is specified only for Medicare older adults, 65 years and older. NCQA is actively engaged with partners including the CMS Office of Minority Health in identifying feasible methods to further integrate social risk factors into health plan quality measures, with a focus on stratification. This is aligned with recent recommendations from MedPAC and ASPE on optimal methods for addressing social risk in quality measurement and programs.1,2This is an NCQA wide initiative. Our intent is to implement methods to bridge data concerns in the future.

1. Medicare Payment Advisory Commission. (2020). The Medicare Advantage program: Status report. In Report to the Congress: Medicare Payment Policy (p. 397). <http://medpac.gov/docs/default-source/reports/mar20_medpac_ch13_sec.pdf>
2. Office of the Assistant Secretary for Planning and Evaluation, & U.S. Department of Health & Human Services. (2020). Second Report to Congress on Social Risk and Medicare’s Value-Based Purchasing Programs. <https://aspe.hhs.gov/social-risk-factors-and-medicares-value-basedpurchasing-programs>

**2016 Submission**

*This question was not on the 2016 form.*

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

**2a2. RELIABILITY TESTING**

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

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

**2020 Submission**

Reliability testing of performance measure score

We utilized the methodology described by John Adams (Adams, J.L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009) to calculate signal-to-noise reliability. This methodology uses the Beta-binomial model to assess how well one can confidently distinguish the performance of one reporting entity from another. Conceptually, the Beta-binomial model is the ratio of signal to noise. The signal is the proportion of the variability in measured performance that can be explained by real differences across reporting entities (plans, physicians, etc.) in performance. The Beta-binomial model is an appropriate model when estimating the reliability of simple pass/fail rate measures, such as the *Use of High-Risk Medications in Older Adults* measure. Reliability scores range from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (i.e., noise), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance across reporting entities.

For the *Use of High-Risk Medications in Older Adults* measure, health plans are the reporting entity. For the formulas and explanations below, we use health plans as the reporting entity.

The formula for signal-to-noise reliability is:

Signal-to-noise reliability = σ2plan-to-plan / (σ2plan-to-plan + σ2error)

More simply, the formula is the numerator is the variation across plans, and the denominator is the sum of the variation across plans plus the variation within the plan (across members).

Therefore, we need to estimate two variances: 1) variance between plans (σ2plan-to-plan); 2) variance within plans (σ2error).

1. Variance between plans = σ2plan-to-plan = (α β) / (α + β + 1)(α + β)2

α and β are two shape parameters of the Beta-Binomial distribution, α >0, β > 0

1. Variance within plans: σ2error = p̂(1- p̂)/n

p̂ = observed rate for the plan

n = plan-specific denominator for the observed rate (most often the number of eligible plan members)

Using Adams’ 2009 methodology, we estimated the reliability for each reporting entity, then averaged these reliability estimates across all reporting entities to produce a point estimate of signal-to-noise reliability. We label this point estimate “mean signal-to-noise reliability”. The mean signal-to-noise reliability measures how well, on average, the measure can differentiate between reporting entity performance on the measure.

Along with the point estimate of mean signal-to-noise reliability, we are also providing:

1. The standard error (SE) and 95% confidence interval (95% CI) of the mean signal-to-noise reliability for all plans and stratified by the denominator size (number of eligible members per plan). The SE and 95% CI of the mean signal-to-noise reliability provides information about the stability of reliability. The 95% CI is the mean signal-to-noise reliability ± (1.96\*SE). The narrower the confidence interval, the less the mean signal-to-noise reliability estimate will change due to idiosyncratic features of specific plans. We also stratified the results by the denominator size using terciles of the distribution to provide additional information about the stability of reliability.
2. The distribution (minimum, 10th, 25th, 50th, 75th, 90th, maximum) of the plan-level signal-to-noise reliability estimates. Each plan’s reliability estimate is a ratio of signal to noise, as described above [ σ2plan-to-plan / (σ2plan-to-plan + σ2error)]. Variability between plans (σ2plan-to-plan) is the same for each plan, while the specific plan error (σ2error) varies. Reliability for each plan is an ordinal measure of how well one can determine where a given plan lies in the distribution of reliability across all plans, with higher estimates indicating better reliability. We also stratified the results by the denominator size using terciles of the distribution to provide additional information about the distribution of plan-level signal-to-noise reliability estimates. The number of plans in each stratum and the per-plan denominators of the performance rates are displayed in the summary tables.

This methodology allows us to estimate the reliability for each plan and summarize the distribution of these estimates.

**2016 Submission**

Reliability was estimated by using the beta-binomial model. Beta-binomial is a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® health plan measures. The beta-binomial model assumes the plan score is a binomial random variable conditional on the plan´s true value that comes from the beta distribution. The beta distribution is usually defined by two parameters, alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The beta distribution can be symmetric, skewed or even U-shaped.

Reliability used here is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance. The higher the reliability score, the greater is the confidence with which one can distinguish the performance of one plan from another. A reliability score greater than or equal to 0.7 is considered very good.

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
**2020 Submission**

Table 2a. Mean Signal-To-Noise Reliability, Standard Error (SE) and 95% Confidence Interval (95% CI) for the *Use of High-Risk Medications in Older Adults* Measure by Terciles of the Denominator Size and for All Submissions, 2018

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stratification | Number of Plans | Number of Eligible Members per Plan (min - max) | Mean Signal-To-Noise Reliability | SE | 95% CI |
| *Use of High-Risk Medications in Older Adults* | **502** | **32 - 679844** | **0.936** | **0.006** | **(0.924, 0.947)** |
| Tercile 1 | 166 | 32 - 2456 | 0.857 | 0.012 | (0.833, 0.881) |
| Tercile 2 | 165 | 2469 - 15564 | 0.986 | 0.001 | (0.985, 0.988) |
| Tercile 3 | 171 | 15856 - 679844 | 0.997 | 0.000 | (0.997, 0.997) |

SE: Standard Error of the mean.

95% CI: 95% confidence interval.

Table 2b. Distribution of Plan-Level Signal-To-Noise Reliability for the *Use of High-Risk Medications in Older Adults* Measure by Terciles of the Denominator Size and for All Submissions, 2018

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Distribution of Plan Estimates of Signal-to-Noise Reliability | | | | | | |
| Stratification | Number of Plans | Min | P10 | P25 | P50 | P75 | P90 | Max |
| *Use of High-Risk Medications in Older Adults* | 502 | 0.193 | 0.798 | 0.950 | 0.988 | 0.998 | 0.999 | 1.000 |
| Tercile 1 | 166 | 0.249 | 0.615 | 0.812 | 0.928 | 0.962 | 0.980 | 1.000 |
| Tercile 2 | 165 | 0.962 | 0.975 | 0.982 | 0.988 | 0.993 | 0.995 | 0.997 |
| Tercile 3 | 171 | 0.987 | 0.994 | 0.996 | 0.998 | 0.999 | 0.999 | 1.000 |

**2016 Submission**

Using 2014 HEDIS Health Plan performance data, reliability for this measure was calculated as 0.99814 for receipt of one or more high-risk prescriptions and 0.99594 for receipt of two or more high-risk prescriptions.

**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?*)  
**2020 Submission**

Table 2a provides the point estimate of mean signal-to-noise reliability, its standard error, and the 95% CI for the *Use of High-Risk Medications in Older Adults* measure stratified by the denominator size (distribution of the number of eligible members per plan). The reliability estimate is 0.936, and the 95% CI is (0.924, 0.947), indicating very good reliability for the measure. Stratified analyses show that reliability increases as plan size gets larger.

Table 2b summarizes the distribution of plan-level signal-to-noise reliability estimates for the *Use of High-Risk Medications in Older Adults* measure. The estimates range from 0.193 to 1.000. The 50th percentile is 0.988, which exceeds the 0.70 threshold for reliability. This table also includes the distribution of plan-level signal-to-noise reliability estimates stratified by the tercile of the denominator size. Note that the low minimum reliability estimate (0.193) and low observed reliability in the first tercile is likely explained by a handful of very small plans (small denominators) who inflate the sigma-squared error in the signal-to-noise calculation (see 2a2.2, above). Very high reliability is observed in a majority of plans and reliability estimates are higher for plans with a larger denominator.

**2016 Submission**

Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (signal). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. The testing suggests that both indicators in this measure are highly reliable.

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

**2b1. VALIDITY TESTING**

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

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*) **NOTE**: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

**2b1.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

**2020 Submission**

Method of testing construct validity

We tested for construct validity by exploring the following:

* Is the *Use of High-Risk Medications in Older Adults* measure correlated with the HEDIS *Potentially Harmful Drug-Disease Interactions in Older Adults* measure, which assesses the percentage of Medicare members 65 years and older who have evidence of an underlying disease, condition or health concern and who were dispensed an ambulatory prescription for a potentially harmful medication, concurrent with or after the diagnosis?

We hypothesized that the *Use of High-Risk Medications in Older Adults* measure would be correlated with the rates of *Potentially Harmful Drug-Disease Interactions in Older Adults* measure, particularly Rate 2 (Dementia) and Rate 3 (Chronic kidney disease). In addition, organizations that perform well on *Use of High-Risk Medications in Older Adults* should perform well on the other medication safety measure, *Potentially Harmful Drug-Disease Interactions in Older Adults*, given that they address the same older adult population.

NCQA performs Pearson correlation for construct validity using HEDIS health plan data. The test estimates the strength of linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a strong positive linear association: an increase in values of one variable is associated with increase in value of another variable. A value of 0 indicates no linear association. A value of -1 indicates a strong negative relationship in which an increase in values of the first variable is associated with a decrease in values of the second variable. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The sample size for the correlation analysis is the number of plans that reported both measures. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We adjusted our p-values to account for testing multiple correlations and used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone.

Systematic Assessment of Face Validity of Performance Measure Score

NCQA develops measures using a standardized process. For new measures, face validity is assessed at various steps as described below.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical measurement advisory panels (MAPs), whose members are authorities on clinical priorities for measurement, participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness, and feasibility. Measures are aligned with clinical guidelines whenever possible; the *Use of High-Risk Medications in Older Adults* measure is aligned with the American Geriatrics Society (AGS) Beers Criteria, which recommends drugs to be avoided in older adults. This information is gathered into a work-up format, which is vetted by the MAPs, including the Geriatric Measurement Advisory Panel (GMAP), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM), as well as other panels as necessary.

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

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA about proposed new measures. Public comment offers an opportunity to assess the validity, feasibility, importance and other attributes of a measure from a wider audience. For this measure, a majority of public comment respondents supported the measure. NCQA MAPs and the technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM.  Face validity is then again systematically assessed by the CPM. The CPM reviews all comments before making a final decision and votes to recommend approval of new measures for HEDIS. NCQA’s Board of Directors then approves new measures.

**2016 Submission**

**Method of Assessing Face Validity:** This measure was tested for face validity with two panels of experts. See Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development for names and affiliations of expert panel members.

* The Geriatric Measurement Advisory Panel (GMAP) included 11 experts in geriatrics, including representation by consumers, health plans, health care providers and policy makers.
* NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 16 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

**Method of Testing Construct Validity:** We tested for construct validity by exploring whether the two rates within this measure were correlated with each other and with another measure of medication safety. We hypothesized that organizations that perform well on one of the indicators should perform well on the other indicator as well as the other medication safety measure. To test these correlations we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
**2020 Submission**

Table 3. Health-Plan Level Pearson Correlation Coefficients Among *Use of High-Risk Medications in Older Adults* and *Potentially Harmful Drug-Disease Interactions in Older Adults* Performance Scores, 2018

|  |  |
| --- | --- |
| **Measure** | **Correlation Coefficient** |
|  | *Use of High-Risk Medications in Older Adults* |
| *Use of High-Risk Medications in Older Adults* |  |
| Drug-disease interaction: History of Falls\* | 0.62 |
| Drug-disease interaction: Dementia\* | 0.53 |
| Drug-disease interaction: Chronic Kidney Disease\* | 0.24 |

Note: All correlations are significant at p<0.001

\*The *Potentially Harmful Drug-Disease Interactions in Older Adults* measure has three rates. The first rate assesses the percentage of patients 65 and older with a history of falls who received a high-risk medication. The second rate assesses the percentage of patients 65 and older with dementia who received a high-risk medication. The third rate assesses the percentage of patients 65 and older with chronic kidney disease who received a high-risk medication.

Results of face validity assessment

NCQA worked closely with our multi-stakeholder MAPs to re-evaluate the measure based on the latest recommendations in the American Geriatric Society’s 2019 Beers Criteria. The last Beers Criteria update prior to this publication was in 2015. Based on the 2019 Beers Criteria, the primary changes to the measure were updates to medications and retirement of rate 1, which focused on one dispensing event for a high-risk medication. The measure changes were evaluated in 2019. After reviewing, the CPM recommended to send the updated measure to public comment with a majority vote in 2019. The measure was released for Public Comment in 2019 prior to publication in HEDIS. Input from advisory panels and the public comment indicate the measure has face validity.

**2016 Submission**

**Results of Face Validity Assessment:** This measure was developed to address high-risk medication use in the elderly. NCQA and the GMAP worked together to assess which medications to include based on recommendations in the AGS Beers Criteria. The measure was field-tested from 2004-2005. After reviewing field test results the CPM recommended to send the measure to public comment with a majority vote in 2006. The measure was released for Public Comment in 2006 prior to publication in HEDIS. The CPM recommended moving this measure to first year data collection by a majority vote. The measure was then introduced in HEDIS 2007. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following year. The CPM recommended moving this measure public reporting with a majority vote. In summary, the measure was deemed to have the desirable attributes of a HEDIS measure in 2006 (relevance, scientific soundness, and feasibility). These results indicate the MAPs and CPM showed agreement that the measures as specified will accurately differentiate quality across health plans. Our interpretation of these results is that this measure has sufficient face validity.

**Results of Construct Validity Testing:** The results in Table 1a indicate that there was a high correlation between the first and second rate in the measure. There were moderate correlations between both rates and the four rates in the other medication safety measure.

|  |  |  |
| --- | --- | --- |
| Table 1a. Correlations among both rates in the measure and a drug-disease interaction measure1 | | |
| Measure | Pearson Correlation Coefficients | |
| Rate 1: One high-risk medication | Rate 2: Two high-risk medications |
| Rate 1: One high-risk medication |  |  |
| Rate 2: Two high-risk medications | .8745 |  |
| Drug-disease interaction: History of Falls | 0.307 | .2735 |
| Drug-disease interaction: Dementia | 0.454 | .4390 |
| Drug-disease interaction: Chronic Kidney Disease | 0.367 | .3552 |
| Drug-disease interaction: Total | 0.386 | .3913 |
| Note: All correlations are significant at p<.05  1The *Potentially Harmful Drug-Disease Interactions in the Elderly* measure has four rates. The first rate assesses the percentage of patients 65 and older with a history of falls who received a high-risk medication. The second rate assesses the percentage of patients 65 and older with dementia who received a high-risk medication. The third rate assesses the percentage of patients 65 and older with chronic kidney disease who received a high-risk medication. The fourth rate is the sum of the three numerators divided by the sum of the three denominators for the three previous rates. Note: “high-risk” medications for each condition are based on recommendations in Table 3 of the American Geriatrics Society Beers Criteria. | | |

**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?*)  
**2020 Submission**

For the purposes of this analysis and the intended use of this measure to evaluate the quality of care for members across health plans, correlation was considered high (strong) if the correlation coefficient is 0.75 to 1, moderate if 0.25 to 0.75, and low (weak) if 0 to 0.25.

Correlations between the *Use of High-Risk Medications in Older Adults* and the *Potentially Harmful Drug-Disease Interactions in Older Adults*measure rates were moderate. This suggests that plans that perform well on the *Use of High-Risk Medications in Older Adults* measure are moderately likely to perform well on the *Potentially Harmful Drug-Disease Interactions in Older Adults* measure. The results indicate that the measure has good validity.

**2016 Submission**

Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.05 to evaluate the test results. P-values less than this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results confirmed the hypothesis that rates in the measure are correlated with each other as well as with another measure of medication safety, suggesting they represent the same underlying quality construct of prescribing inappropriate medications for patients with the corresponding illnesses. These results indicate the measure is a valid measure of a plan’s quality at managing use of high-risk medications in the elderly.

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

**2b2. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b3***](#section2b4)

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

**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*)

**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*)

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

**2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b4***](#section2b5)***.***

**2020 Submission**

N/A. Not an intermediate or health outcome, PRO-PM, or resource use measure.

**2016 Submission**

**2b3.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** Click here to enter number of factors **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

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

**2b3.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

**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?

**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?**

**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*** [***2b3.9***](#question2b49)

**2b3.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b3.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b3.9. Results of Risk Stratification Analysis**:

**2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

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

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

**2b4.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*   
**2020 Submission**

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure.

To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the below 25th and above 75th percentile groups. The t-test method calculates a testing statistic based on the sample size, performance rate, and standard error of each plan. The test statistic is then compared against a t distribution, which is similar to a normal distribution. If the p-value of the test statistic is less than .05, then the two plans’ performance is significantly different from each other.

**2016 Submission**

Comparison of means and percentiles; analysis of variance against established benchmarks: if sample size is >400, we would use an analysis of variance.

**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*)  
**2020 Submission**

Table 4. Variation in Performance, 2018

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Measure | N | Mean eligible population | Mean  rate (%) | SD | Min | P10 | P25 | P50 | P75 | P90 | Max | IQR | p-value |
| 2018 | *Use of High-Risk Medications in Older Adults* | 502 | 28,463 | 9.6 | 3.9 | 0.0 | 5.8 | 7.1 | 8.6 | 11.4 | 14.9 | 27.3 | 4.3 | p <0.001 |

*N = Number of plans reporting*

*IQR = Interquartile range*

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

**2016 Submission**

**2012 to 2014 HEDIS Health Plan Performance Data**

At least one high-risk prescription

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Number of Plans | Mean | Standard Deviation | Min | Max | 10th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 90th Percentile |
| 2012 | 498 | 21.0 | 6.4 | 5.5 | 54.6 | 14.0 | 16.5 | 19.9 | 24.5 | 30.0 |
| 2013 | 494 | 18.0 | 6.1 | 1.0 | 50.5 | 11.5 | 13.8 | 16.7 | 21.1 | 25.8 |
| 2014 | 488 | 13.2 | 6.0 | 2.6 | 46.8 | 7.6 | 9.2 | 11.6 | 16.1 | 21.7 |

At least two high-risk prescriptions

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Number of Plans | Mean | Standard Deviation | Min | Max | 10th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 90th Percentile |
| 2012 | 498 | 6.5 | 2.9 | 1.2 | 25.2 | 3.5 | 4.7 | 6.0 | 7.8 | 10.1 |
| 2013 | 494 | 3.1 | 2.3 | 0.0 | 20.6 | 1.1 | 1.7 | 2.4 | 4.0 | 6.0 |
| 2014 | 488 | 2.1 | 2.0 | 0.0 | 20.8 | 0.6 | 0.9 | 1.4 | 2.5 | 4.6 |

**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?*)  
**2020 Submission**

There is a 4.3 percentage point gap in performance between Medicare plans at the 25th and 75th percentiles. This gap represents on average 1,138 more older adults with at least two high-risk medications in low performing Medicare plans compared to high performing plans. The difference in performance between plans in the 25th percentile and 75th percentile is statistically significant.

**2016 Submission**

*This question was not on the 2016 form.*

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

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

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

**2020 Submission**

This measure has only one set of specifications.

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

**2b5.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)

**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*)

**2b5.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)

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

**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*)  
**2020 Submission**

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

The HEDIS Compliance Audit addresses the following functions:

* Information practices and control procedures
* Sampling methods and procedures
* Data integrity
* Compliance with HEDIS specifications
* Analytic file production
* Reporting and documentation

**2016 Submission**

*This question was not on the 2016 form.*

**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*)  
**2020 Submission**

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

**2016 Submission**

*This question was not on the 2016 form.*

**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*)

**2020 Submission**

All health plans that reported 2018 HEDIS data for this measure reported valid rates as determined by NCQA-certified auditors. This means that auditors did not find any missing data sources for any of the health plan data submissions and determined that none of the rates were materially biased.

**2016 Submission**

*This question was not on the 2016 form.*