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

**Measure Number** (*if previously endorsed*)**:** 0004

**Measure Title**: Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment

**Date of Submission**: 8/15/2018

**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** 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 decision making) 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 non-responders) 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*).

**2018 submission**

**N/A**

**2012 Submission**

This measure is based on administrative claims collected in the course of providing care to health plan members. NCQA collects the Healthcare Effectiveness Data and Information Set (HEDIS) data for this measure directly from health plans via the Interactive Data Submission System (IDSS) portal.

The URL is: <http://www.ncqa.org/tabid/370/default.aspx>

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

**2018 Submission**

Testing of measure score reliability and construct validity was performed using data from 2017.

2012 submission: 2010 data

**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*)

**2018 Submission**

Data for measure score reliability testing: The measure score reliability was calculated from data in calendar year 2017. The number of health plans in the sample for the Initiation indicator included 408 Medicare health plans, 186 Medicaid health plans, and 384 commercial health plans. The number of health plans in the sample for the Engagement indicator included 408 Medicare health plans, 188 Medicaid health plans, and 384 commercial health plans. The sample data included all Medicare, Medicaid and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Systematic evaluation of face validity: NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers, and policy makers. This panel is made up of 21 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

Data for Construct Validity Testing: Construct validity was calculated from HEDIS 2018 data, which represents calendar year 2017. The number of health plans in the sample for the Initiation indicator included 408 Medicare health plans, 186 Medicaid health plans, and 384 commercial health plans. The number of health plans in the sample for the Engagement indicator included 108 Medicare health plans, 188 Medicaid health plans, and 384 commercial health plans. The sample data included all Medicare, Medicaid and commercial health plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

**2012 Submission**

HEDIS Health Plan performance data for the 2010

**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*)   
**2018 Submission**

Patient population for measure score reliability testing: The most recent available data indicates that for 2016, HEDIS data covered 111.5 million commercial health plan members, 53.4 million Medicaid members and 19 million Medicare beneficiaries. Data is summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the population measured for 2017. It includes number of health plans included in the analysis and the median eligible population for the measure across health plans.

|  |  |  |  |
| --- | --- | --- | --- |
| Product Line | Measure Indicator | Number of Plans | Median number of eligible patients for this measure per plan |
| Commercial | Initiation | 384 | 1,617 |
| Engagement | 384 | 1,617 |
| Medicare | Initiation | 408 | 1,328 |
| Engagement | 408 | 1,328 |
| Medicaid | Initiation | 186 | 3,967 |
| Engagement | 188 | 3,963 |

Patient population for Construct Validity Testing: The most recent available data indicates that for 2016, HEDIS data covered 111.5 million commercial health plan members, 53.4 million Medicaid members and 19 million Medicare beneficiaries. Data is summarized at the health plan level. Data are stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the measured entities that include HEDIS data collection and the median eligible population for the measure across health plans for 2017.

|  |  |  |  |
| --- | --- | --- | --- |
| Product Line | Measure Indicator | Number of Plans | Median number of eligible patients for this measure per plan |
| Commercial | Initiation | 384 | 1,617 |
| Engagement | 384 | 1,617 |
| Medicare | Initiation | 408 | 1,328 |
| Engagement | 408 | 1,328 |
| Medicaid | Initiation | 186 | 3,967 |
| Engagement | 188 | 3,963 |

**2012 Submission**

HEDIS Health Plan performance data for the 2010

**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**.

**2018 Submission**

Reliability of the measure score was tested using a beta-binomial calculation. This analysis included the entire HEDIS data sample (described above).

Validity was demonstrated through a systematic assessment of face validity and construct validity. Per NQF instructions, we have described the composition of the technical expert panel which assessed face validity of the measure. Construct validity was demonstrated through a correlation analysis.

**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.

**2018 Submission**

We did not analyze social risk factors. Measure performance was assessed by Medicaid, commercial and Medicare plan types, which serves as a proxy for income.

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

**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*)

**2018 Submission**

Reliability Testing of Performance Measure Score: We utilized the Beta-binomial model (Adams 2009) to assess how well one can confidently distinguish the performance of one accountable 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 in performance. The Beta-binomial model is an appropriate model when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS measures. 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 accountable entities).

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

**2012 Submission**

In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009) in work produced for the National Committee for Quality Assurance (NCQA).

The following is quoted from the tutorial which focused on provider-level assessment: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.”  This approach is also relevant to health plans and other accountable entities.

The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities.  Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance), whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Adams’ approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures.

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

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
**2018 Submission**

Beta-binomial reliability

|  |  |  |  |
| --- | --- | --- | --- |
| Measure Indicator Rate | Beta Binomial Reliability | | |
| Commercial Product | Medicare Product | Medicaid Product |
| Initiation | 0.97 | 0.99 | 0.99 |
| Engagement | 0.94 | 0.96 | 0.99 |

**2012 Submission**

Initiation of AOD Treatment

Commercial

Total: 0.962184

13 – 17 Years: 0.697888

18 Years and Older: 0.961216

Medicaid

Total: 0.9836665

13 – 17 Years: 0.930377

18 Years and Older: 0.983049

Medicare

Total: 0.9732890

18 Years and Older: 0.9732890

Engagement of AOD Treatment.

Total: 0.967456

13 – 17 Years: 0.788911

18 Years and Older: 0.965894

Medicaid

Total: 0.992259

13 – 17 Years: 0.961001

18 Years and Older: 0.99193

Medicare

Total: 0.872810

18 Years and Older: 0.872810

**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?*)  
**2018 Submission**

Interpretation of measure score reliability testing:

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

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

**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)*  
**2018 submission:**

Method of testing construct validity: We tested for construct validity by exploring whether *Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment* (IET) was correlated with the *Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence* measure(FUA). We also examined whether the two indicators within the Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment measure were correlated with each other. We hypothesized that organizations that perform well on the FUA measure should also perform well on the IET measure given that they are similar concepts. We also hypothesized that health plans perform well on one of the two indicators in the Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment should perform well on the other indicator because they are similar constructs.

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 to +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. 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.

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, as p-values less than this threshold imply it is unlikely that a non-zero coefficient was observed due to chance alone.

For this measure, we specifically hypothesized:

1. Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence (both 7 Day and 30 day follow-up indicators) will be positively correlated with Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment (both initiation and engagement indicators) (i.e. plans that have high rates of follow-up will have high rates of inhiation and engagement in treatment)
2. The Initiation and Engagement of Alcohol and Other Drug Abuse and Dependence Treatment Initiation Rate will be positively correlated with the Engagement Rate (i.e. plans that have high rates of initiation of treatment will have high rates of engagement in treatment).

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

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert 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. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

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

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

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

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

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

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

**2012 Submission**

NCQA identified and refined measure management into a standardized process called the HEDIS measure life cycle.

\*Step 1: Topic selection is the process of identifying measures that meet criteria consistent with the overall model for performance measurement. There is a huge universe of potential performance measures for future versions of HEDIS. The first step is identifying measures that meet formal criteria for further development.

NCQA staff identifies areas of interest or gaps in care. Clinical expert 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. This information is gathered into a work-up format. Refer to What Makes a Measure “Desirable”? The work-up is vetted by NCQA’s MAPs, the TAG, the HEDIS Policy Panel and various other panels.

\*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.Ensure funding throughout measure testing

2.Prepare a detailed conceptual and operational work-up that includes a testing proposal

3.Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures

The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

\*Step 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to the CPM about new measures or about changes to existing measures.

NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

\*Step 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA’s Quality Compass? or in accreditation scoring.

The first-year distinction guarantees that a measure can be efficiently collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA’s experience is that the first year of large-scale data collection often reveals unanticipated issues.

After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

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

Step 6: Evaluation is the ongoing review of a measure’s performance and recommendations for its modification or retirement. Every measure is reevaluated at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments contribute to measure evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

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

What makes a measure “Desirable”?

Whether considering the value of a new measure or the continuing worth of an existing one, we must define what makes a measure useful. HEDIS measures encourage improvement. The defining question for all performance measurement—”Where can measurement make a difference?”—can be answered only after considering many factors. NCQA has established three areas of desirable characteristics for HEDIS measures, discussed below.

1. Relevance: Measures should address features that apply to purchasers or consumers, or which will stimulate internal efforts toward quality improvement. More specifically, relevance includes the following attributes.

Meaningful: What is the significance of the measure to the different groups concerned with health care? Is the measure easily interpreted? Are the results meaningful to target audiences?

Measures should be meaningful to at least one HEDIS audience (e.g., individual consumers, purchasers or health care systems). Decision makers should be able to understand a measure’s clinical and economic significance.

Important to health: What is the prevalence and overall impact of the condition in the U.S. population? What significant health care aspects will the measure address?

We should consider the type of measure (e.g., outcome or process), the prevalence of medical condition addressed by the measure and the seriousness of affected health outcomes.

Financially important: What financial implications result from actions evaluated by the measure? Does the measure relate to activities with high financial impact?

Measures should relate to activities that have high financial impact.

Cost effective: What is the cost benefit of implementing the change in the health care system? Does the measure encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness? Measures should encourage the use of cost-effective activities or discourage the use of activities that have low cost-effectiveness.

Strategically important: What are the policy implications? Does the measure encourage activities that use resources efficiently? Measures should encourage activities that use resources most efficiently to maximize member health.

Controllable: What impact can the organization have on the condition or disease? What impact can the organization have on the measure? Health care systems should be able to improve their performance. For outcome measures, at least one process should be controlled and have an important effect on outcome. For process measures, there should be a strong link between the process and desired outcome.

Variation across systems: Will there be variation across systems? There should be the potential for wide variation across systems.

Potential for improvement: Will organizations be able to improve performance? There should be substantial room for performance improvement.

2. Scientific soundness: Perhaps in no other industry is scientific soundness as important as in health care. Scientific soundness must be a core value of our health care system—a system that has extended and improved the lives of countless individuals.

Clinical evidence: Is there strong evidence to support the measure? Are there published guidelines for the condition? Do the guidelines discuss aspects of the measure? Does evidence document a link between clinical processes and outcomes addressed by the measure? There should be evidence documenting a link between clinical processes and outcomes.

Reproducible: Are results consistent? Measures should produce the same results when repeated in the same population and setting.

Valid: Does the measure make sense? Measures should make sense logically and clinically, and should correlate well with other measures of the same aspects of care.

Accurate: How well does the measure evaluate what is happening? Measures should precisely evaluate what is actually happening.

Risk adjustment: Is it appropriate to stratify the measure by age or another variable? Measure variables should not differ appreciably beyond the health care system’s control, or variables should be known and measurable. Risk stratification or a validated model for calculating an adjusted result can be used for measures with confounding variables.

Comparability of data sources: How do different systems affect accuracy, reproducibility and validity? Accuracy, reproducibility and validity should not be affected if different systems use different data sources for a measure.

3. Feasibility:

The goal is not only to include feasible measures, but also to catalyze a process whereby relevant measures can be made feasible.

Precise specifications: Are there clear specifications for data sources and methods for data collection and reporting? Measures should have clear specifications for data sources and methods for data collection and reporting.

Reasonable cost: Does the measure impose a burden on health care systems? Measures should not impose an inappropriate burden on health care systems.

Confidentiality: Does data collection meet accepted standards of member confidentiality?

Data collection should not violate accepted standards of member confidentiality. Logistical feasibility

Are the required data available?

Auditability: Is the measure susceptible to exploitation or “gaming” that would be undetectable in an audit? Measures should not be susceptible to manipulation that would be undetectable in an audit.

**2b1.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)  
**2018 Submission**

Results of construct validity testing:

The results in Table 1a describe the correlations observed for Commercial plans. The results indicate that the Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment *Initiation indicator* and the Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence 7 and 30 Day Indicators had significant weak positive correlations (0.19 and 0.16, respectively). The Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment *Engagement indicator* and the Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence 7 and 30 Day Indicators had significant moderate positive correlations (0.31 and 0.31, respectively).

The results in Table 1b describe the correlations observed for Medicare plans. The results indicate that the Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment Initiation indicator and the Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence 7 and 30 Day Indicators had significant weak positive correlations (0.24 and 0.26, respectively). The Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment Engagement indicator and the Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence 7 and 30 Day Indicators had significant moderate positive correlations (0.39 and 0.41, respectively).

The results in Table 1c describe the correlations observed for Medicaid plans. The results indicate that the Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment Initiation indicator and the Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence 7 and 30 Day Indicators had insignificant correlations (0.13 and 0.08, respectively). The Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment Engagement indicator and the Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence 7 and 30 Day Indicators had significant moderate positive correlations (0.57 and 0.60, respectively).

The results in Tables 1a, 1b, and 1c also indicate that the Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment indicators were significantly (p<.05) correlated with each other in the direction that was hypothesized (positively). The level of correlations among these indicators is moderate (0.51- 0.59) across the various product lines (Medicare, Medicaid, commercial).

**Table 1a. Correlation between Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment and Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence in Commercial Plans – HEDIS 2018**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure/Measure Element** | **Pearson Correlation Coefficients** | | | |
| **FUA: 7 Day Indicator** | **FUA: 30 Day Indicator** | **IET: Initiation Indicator** | **IET: Engagement Indicator** |
| **IET: Initiation Indicator** | 0.19  P value: 0.0008 | 0.16  P value: 0.005 | 1 | 0.51  P value: <.0001 |
| **IET: Engagement Indicator** | 0.31  P value: <.0001 | 0.31  P value: <.0001 | 0.51  P value: <.0001 | 1 |

IET: Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment

FUA: Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence

**Table 1b. Correlations between Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment and Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence in Medicare Plans – HEDIS 2018**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure/Measure Element** | **Pearson Correlation Coefficients** | | | |
| **FUA: 7 Day Indicator** | **FUA: 30 Day Indicator** | **IET: Initiation Indicator** | **IET: Engagement Indicator** |
| **IET: Initiation Indicator** | 0.24  P value: .0001 | 0.26  P value: <.0001 | 1 | 0.59  P value: <.0001 |
| **IET: Engagement Indicator** | 0.39  P value: <.0001 | 0.41  P value: <.0001 | 0.59  P value: <.0001 | 1 |

IET: Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment

FUA: Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence

**Table 1c. Correlations between Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment and Follow-Up After Emergency Department Visit for Alcohol and other Drug Abuse or Dependence in Medicaid Plans – HEDIS 2018**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure/Measure Element** | **Pearson Correlation Coefficients** | | | |
| **FUA: 7 Day Indicator** | **FUA: 30 Day Indicator** | **IET: Initiation Indicator** | **IET: Engagement Indicator** |
| **IET: Initiation Indicator** | 0.13  P value: 0.10 | 0.08  P value: .31 | 1 | 0.56  P value: <.0001 |
| **IET: Engagement Indicator** | 0.57  P value: <.0001 | 0.60  P value: <.0001 | 0.56  P value: <.0001 | 1 |

IET: Initiation and Engagement of Alcohol and Other Drug Abuse or Dependence Treatment

FUA: Follow-Up After Emergency Department Visit for Alcohol and Other Drug Abuse or Dependence

Results of face validity assessment:

Since the last endorsement of this measure, small updates were made to bring the measure into alignment with the most recent clinical practice guidelines and to improve the face validity of the measure. These updates include the inclusion of pharmacotherapy for the treatment of opioid and alcohol abuse and dependence and the inclusion of telehealth as an appropriate way to deliver treatment for those with substance abuse and dependence. Results from multiple multi-stakeholder measurement advisory panels, as well as those submitting to public comment, indicate that the measure as specified will accurately differentiate quality across providers and has sufficient face validity.

**2012 Submission**

Step 1: The Initiation and Engagement of Alcohol and Other Drug Dependence measure was developed to address a gap in care concerning follow-up care for people with alcohol or other drug dependence. NCQA’s Performance Measurement Department, the Behavioral Health MAP and The Washington Circle worked together to assess the most appropriate tools for monitoring follow-up for AOD.

Step 2: The measure was written, field-tested, and presented to the CPM in 2004. The CPM recommended to send the measure to public comment with a vote of 14 in favor and none opposed.

Step 3: The measure was released for Public Comment in spring 2004. We received and responded to comments on this measure. The CPM recommended moving this measure to first year data collection with a vote of 14 in favor and none opposed.

Step 4: The Initiation and Engagement of Alcohol and Other Drug Dependence measure was introduced in HEDIS 2005. 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 vote of 16 in favor and none opposed.

Step 5: The Initiation and Engagement of Alcohol and Other Drug Dependence measure was reevaluated in 2011/2012.

**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?*)  
**2018 Submission**

Results of face validity assessment:

Results from multiple multi-stakeholder measurement advisory panels, as well as those submitting to public comment, indicate that the measure as specified will accurately differentiate quality across providers and has sufficient face validity.

*Interpretation of construct validity testing:* The results confirmed the hypothesis that health plans with high rates of follow-up also have high rates of initiation and engagement in treatment (exception seen in the Medicaid population; only engagement in treatment had significant positive correlation with follow-up). The results also confirmed the hypothesis that the Initiation and Engagement measure indicators are correlated with each other, suggesting they represent the same underlying quality construct of substance abuse and dependence care. These results indicate that the Initiation and Engagement measure is a valid measure of a plan’s quality of managing substance abuse or dependence treatment.

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

**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*)  
Testing was not performed for exclusions.

**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*)  
Testing was not performed for exclusions.

**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*)  
Testing was not performed for exclusions.

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

**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.***

**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***

**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)*

**2018 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 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than 0.05, then the two plans’ performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measures entities. However, the method can be used for comparison of any two measured entities

**2012 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*)  
**2018 Submission**

Variation in Performance across Health Plans for Initiation Indicator in 2017 Data

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Commercial | 1,617 | 36.7 | 7.7 | 29.4 | 33.0 | 35.9 | 39.2 | 42.0 | 6.2 | 0.0250 |
| Medicare | 1,328 | 34.4 | 13.1 | 15.2 | 27.0 | 35.1 | 41.7 | 48.3 | 14.7 | <0.001 |
| Medicaid | 3,967 | 42.3 | 7.4 | 33.7 | 38.6 | 42.2 | 46.4 | 50.2 | 7.8 | <0.001 |

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

IQR: Interquartile range

P-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile. P-values are less than 0.05.

Variation in Performance across Health Plans for Engagement Indicator in 2017 Data

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Commercial | 1,617 | 13.4 | 4.1 | 8.7 | 11.0 | 13.3 | 15.9 | 18.3 | 4.9 | <0.001 |
| Medicare | 1,328 | 4.2 | 2.9 | 0.8 | 2.2 | 3.7 | 5.6 | 8.1 | 3.4 | .0064 |
| Medicaid | 3,963 | 13.6 | 5.9 | 6.1 | 9.1 | 13.7 | 17.7 | 21.4 | 8.6 | <0.001 |

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

IQR: Interquartile range

P-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile. P-values are less than 0.05.

**2012 submission**

Initiation:

Medicaid

Measurement Year: 2009; 2010; 2011

AVE: 44.52 44.35 42.93

N: 61 68 79

Min: 17.74 22.72 23.86

Max: 69.09 76.71 78.88

SD: 10.01 10.31 10.96

P10: 32.74 31.78 30

P25 37.21 38.42 35.68

P50 43.79 43.92 40.81

P75 51.26 48.79 48.84

P90 57.33 57.31 60.72

Medicare

Measurement Year: 2009; 2010; 2011

AVE: 46.55 48.89 48.02

N: 268 306 368

Min: 5.19 12.12 7.32

Max: 84.85 98.23 98.24

SD: 13.94 15.72 17.07

P10: 29.25 27.41 27.31

P25: 38.48 38.88 36.17

P50: 46.83 49.17 46.08

P75: 54.7 56.9 57.62

P90: 64.29 70.27 74.11

Commercial

Measurement Year: 2009; 2010; 2011

AVE: 42.46 42.28 41.89

N: 415 402 392

Min: 14.71 12.9 16.67

Max: 70.18 72.65 69.77

SD: 7.4 7.32 7.51

P10: 33.47 34.03 33.01

P25: 38.6 38.2 37.42

P50: 42.2 41.79 41.81

P75: 46.67 46.27 45.71

P90: 51.33 50.6 50.27

Engagement:

Medicaid

Measurement Year: 2009; 2010; 2011

AVE: 12.43 12.31 14.19

N: 61 68 79

Min: 0 0.99 0.5

Max: 55.57 54.26 41.44

SD: 11.45 10.73 9.79

P10: 1.69 2.34 2.02

P25: 3.46 4.15 5.72

P50: 10.06 10.18 14.53

P75: 16.79 17.6 20.52

P90: 21.7 21.42 25.89

Medicare

Measurement Year: 2009; 2010; 2011

AVE: 5.36 4.51 4.02

N: 268 311 366

Min: 0 0 0

Max: 41.79 35.64 26.25

SD: 6.23 4.17 3.46

P10: 0.7 0.8 0.56

P25: 1.97 2.08 1.71

P50: 3.13 3.52 3.19

P75: 6.32 5.78 5.61

P90: 11.63 8.53 7.95

Commercial

Measurement Year: 2009; 2010; 2011

AVE: 16.2 15.93 15.78

N: 415 402 392

Min: 0 1.61 0.85

Max: 53.4 46.99 46.45

SD: 5.7 5.88 5.6

P10: 9.74 8.51 9.54

P25: 12.43 12.19 12.01

P50: 15.85 15.61 15.56

P75: 19.82 19.19 18.68

P90: 22.46 22.19 22.09

**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?*)  
**2018 Submission**

The results above indicate there is a 6.2-14.7% gap in performance for the initiation indicator, and a 3.4-8.6% gap in performance for the engagement indicator between plans performing at the 25th and 75th percentiles. The difference between plan performance at the 25th and 75th percentile is statistically significant for both indicator rates across all product lines.

In commercial plans, there is a 6.2 percentage point gap between 25th and 75th percentile plans for the initiation of treatment indicator rate. This gap represents an average 97 more patients who have initiated treatment for alcohol or other drug abuse or dependence within 14 days of their new diagnosis in high performing plans compared to low performing plans (estimated from average health plan eligible population). For the engagement in treatment indicator rate, there is a 4.9 percentage point gap between 25th and 75th percentile commercial plans. This gap represents an average 79 more patients who have engaged in treatment within the 34 days following initiation of treatment for alcohol or other drug abuse or dependence in high performing plans compared to low performing plans (estimated from average health plan eligible population).

In Medicare plans, there is a 14.7 percentage point gap between 25th and 75th percentile plans for the initiation of treatment indicator rate. This gap represents an average 195 more patients who have initiated treatment for alcohol or other drug abuse or dependence within 14 days of their new diagnosis in high performing plans compared to low performing plans (estimated from average health plan eligible population). For the engagement in treatment indicator rate, there is a 3.4 percentage point gap between 25th and 75th percentile Medicare plans. This gap represents an average 45 more patients who have engaged in treatment within the 34 days following initiation of treatment for alcohol or other drug abuse or dependence in high performing plans compared to low performing plans (estimated from average health plan eligible population).

In Medicaid plans, there is a 7.8 percentage point gap between 25th and 75th percentile plans for the initiation of treatment indicator rate. This gap represents an average 309 more patients who have initiated treatment for alcohol or other drug abuse or dependence within 14 days of their new diagnosis in high performing plans compared to low performing plans (estimated from average health plan eligible population). For the engagement in treatment indicator rate, there is an 8.6 percentage point gap between 25th and 75th percentile Medicaid plans. This gap represents an average 341 more patients who have engaged in treatment within the 34 days following initiation of treatment for alcohol or other drug abuse or dependence in high performing plans compared to low performing plans (estimated from average health plan eligible population).

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

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

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

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

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

**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*)

**2018 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, as applicable:

- Information practices and control procedures

- Sampling methods and procedures

- Data integrity

- Compliance with HEDIS specifications

- Analytic file production

- Reporting and documentation

**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*)  
**2018 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 in any analyses. Once measures, new or re-evaluated, are added to HEDIS, NCQA conducts an analysis to assess the measure’s feasibility for implementation 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.

**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*)

**2018 Submission**

Plans collect this measure using all administrative data sources. NCQA’s audit process checks that plans’ measure calculations are not biased due to missing data.