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

**Measure Number** (*if previously endorsed*)**:** Click here to enter NQF number

**Measure Title**: Prescription of HIV Antiretroviral Therapy

**Date of Submission**: 12/16/2016

**Type of Measure:**

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

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). * For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment. |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.  **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and sociodemographic 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.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions  **15.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

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

**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*).

This measure is a legacy electronic clinical quality measure (eCQM) – an NQF endorsed measure currently used in federal quality programs that has been respecified into eMeasure. Per NQF modified testing requirements for legacy eCQMs, the measure was tested in the Bonnie testing tool. Bonnie is designed to validate eCQM specifications (HQMF output and value sets) against the measure’s expected behavior for user-developed synthetic test patients.

The synthetic patient bundle used to test this measure was designed to simulate clinically relevant, realistic patient scenarios aligned with the target population for this measure. Full details on the Bonnie synthetic patient bundle used to test this measure are included in the Bonnie testing attachment.

For more information on Bonnie, please visit https://bonnie.healthit.gov/.

**1.3. What are the dates of the data used in testing**? The Bonnie test environment simulates the year 2012 as the measurement period.

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

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

**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*)

Not applicable. The Bonnie synthetic patient bundle was used to test the measure.

**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*)

A test bundle of 34 patients was designed and built within the Bonnie testing tool to evaluate the measure logic. Information documented for each patient within the bundle include:

Patient name

Date of birth

Race

Ethnicity

Gender

Payer

Additional elements contained within the patient profiles as appropriate for testing against expected outcomes include:

Diagnosis

Medication orders

Encounters

The patient bundle’s demographics were designed to mimic the HIV/AIDS population, specifically drawing from the patient characteristics collected via the Ryan White HIV/AIDS Program Services Report (RSR).

The breakdown of test bundle demographics for the 34 patients included (represented by number of patients/percentage of bundle): males 23/68%; females 11/32%; American Indian/Alaska Native 1/3%; Asian 1/3%; Black/African American 15/44%; Native Hawaiian/Pacific Islander 0/0%; White 9/26%; Hispanic/Latino 8/24%; younger than 13 1/3%; 13-17 years old 1/3%; 18-24 years old 2/6%; 25-34 years old 6/18%; 35-44 years old 6/18%; 45-54 years old 10/29%; 55-65 years old 6/18%; older than 65 2/6%.

Full details on the Bonnie synthetic patient bundle used to test this measure, including human-readable and QRDA Category 1 format documents for each synthetic patient record, are included in the Bonnie testing attachment.

**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**.

The Bonnie patient test deck was used to satisfy all testing requirements for this measure. The testing results are further supported by testing data for the chart-abstracted version of this measure collected through the Health Resources and Services Administration HIV/AIDs Bureau’s Ryan White HIV/AIDS Program Services Report.

**1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).**

Patient sociodemographic variables considered in the analysis of the chart-abstracted version of this measure were included in the eCQM specifications and modeled in the Bonnie patient bundle. These variables included age, race, ethnicity, gender and payer.

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

**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*)

Currently, there is no performance data available to test the eCQM. However, the chart-abstracted version of this measure has been in use in national quality reporting programs since as early as 2010.

The most recent reliability analysis of the chart-abstracted measure was calculated according to the methods outlined in a technical report prepared by J.L. Adams for the National Committee for Quality Assurance titled “The Reliability of Provider Profiling: A Tutorial” (RAND Corporation, TR-653-NCQA, 2009). In this context, reliability represents the ability of a measure to confidently distinguish the performance of one physician from another. As discussed in the report: “Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of variability in measured performance that can be explained by real differences in performance. There are 3 main drivers of reliability; sample size, differences between physicians, and measurement error.”

According to this approach, reliability is estimated with a beta-binomial model. The beta-binomial model is appropriate for measuring the reliability of pass/fail measures such as those proposed here. Reliability scores vary from 0.0 to 1.0, with a score of zero indicating that all variation is attributable to measurement error (noise, or individual accountable entity variance) whereas a reliability of 1.0 implies that all variation is caused by real difference in performance across accountable entities.

**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*)

Overall reliability scores (i.e., median of provider-level reliability [R\_median], minimum [R\_min], maximum [R\_max]) by year, and the overall variance between sites, are summarized below.

Table 1. Overall reliability scores by year, 2010-2014

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | % suppressed | Var\_between | R\_median | R\_min | R\_max |
| 2010 | 68.4% | 0.069 | 0.990 | 0.354 | 1.000 |
| 2011 | 71.1% | 0.066 | 0.991 | 0.347 | 1.000 |
| 2012 | 74.3% | 0.059 | 0.991 | 0.322 | 1.000 |
| 2013 | 77.5% | 0.048 | 0.991 | 0.276 | 1.000 |
| 2014 | 77.6% | 0.073 | 0.996 | 0.368 | 1.000 |

Reliability varied across providers by year. The proportion of providers with reliability greater than or equal to 0.9, 0.8, and 0.7 are shown below.

Table 2. Distribution of provider-level reliability scores by year, 2010-2014

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | N | ≥0.9  n (%) | ≥0.8  n (%) | ≥0.7  n (%) |
| 2010 | 846 | 793 (93.7) | 819 (96.8) | 836 (98.8) |
| 2011 | 811 | 752 (92.7) | 788 (97.2) | 792 (97.7) |
| 2012 | 816 | 753 (92.3) | 788 (96.6) | 801 (98.2) |
| 2013 | 823 | 753 (91.5) | 794 (96.5) | 806 (97.9) |
| 2014 | 813 | 771 (94.8) | 794 (97.7) | 802 (98.7) |

**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?*)

There is no established cut-off for minimum reliability level. Values above 0.7 are considered sufficient to see differences between providers and the mean, and values above 0.9 are considered sufficient to see differences between pairs of providers (RAND Corporation, TR-653-NCQA, 2009).

Each year, more than 91% of providers had reliability scores of 0.9 or greater. Therefore, the reliability of viral supression can be considered to be sufficient to identify real differences in performance across providers. As previously mentioned, sample size is another driver of reliability and likely contributed to the lowest reliability scores (e.g., in 2014 site 2081 had a reliability of 0.368, and reported 1 of 2 had been prescribed ART). However, median reliability was consistently 0.99 during 2010-2014, supporting the conclusion that the reliability of this measure can be considered very good.

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

**2b2. VALIDITY TESTING**

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

**Performance measure score**

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

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The Bonnie testing environment was used to test the validity of the measure logic and data elements. For each Bonnie synthetic patient, an expected measure result was assigned to reflect the expected outcome of the measure given the specific patient scenario and associated data. The synthetic patients were run against the HQMF output loaded into Bonnie, which produces a measure outcome for each patient and evaluates it against the expected outcome. A patient is considered to pass Bonnie testing when the expected outcome matches the actual outcome, e.g. when a patient is expected to be in the numerator population and the computation of the synthetic patient data against the eCQM logic places the patient in the numerator. In order to achieve a rigorous, clinically relevant test bundle, synthetic patients were designed following the below principles and test areas:

* Clinical relevance. References cited within the chart abstracted measure specification were used to design clinically relevant, realistic patient profiles for the measure’s target population. This approach ensured the eCQM logic maintained alignment with the clinical intent of the chart abstracted measure.
* 100% logic coverage: The resulting bundle of synthetic patients collectively includes all data elements and conditions logic that are specified within the measure logic, including at least one patient evaluating against each measure population pathway. Fully testing the measure logic increases test rigor and mitigates risk of unexpected outcomes.
* Edge case testing. Edge cases refer to those data elements that test the upper or lower boundary of measure logic conditions, e.g. a diagnosis starting on the latest qualifying date or a medication order for antiretroviral therapy starting on the first day or last day of the measurement period. Edge cases are designed to test each edge that exists within each measure population.
* Negative testing. Negative testing involves use of test cases do not evaluate positively against measure logic, but are otherwise clinically relevant and realistic, e.g. scenarios where an HIV diagnosis was not documented or a medication order for antiretroviral therapy starting on the first day or last day of the measurement period. Negative testing further validates measure logic by accurately evaluating patients against expected outcomes and simulating the effect of missing data on measure results.

In addition to Bonnie testing, the measure specifications were reviewed independently by three eCQM experts to confirm the logic was syntactically correct, using appropriate and current versions of the eCQM standards and terminologies, and consistent with the intent of the chart-abstracted measure.

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

Bonnie testing results provide logic coverage and passing rates. The synthetic bundle reached 100% coverage, confirming each logic pathway was tested. The results also showed 100% passing rate, confirming all synthetic patients performed as expected.

Full details on Bonnie testing results are contained in the Bonnie testing attachment. The attachment includes a human-readable (HTML) summary document that lists each patient within the bundle and its passing status against expected measure outcomes. The attachment also includes a summary spreadsheet for the synthetic patient bundle which lists each patient, associated demographics, expected and actual measure population outcomes, and which portions or each measure population logic the patient meets expectations for.

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

The results of measure logic testing through use of Bonnie provided confidence in the measure logic accurately representing the clinical intent and alignment with the chart abstracted measure.

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

**2b3. EXCLUSIONS ANALYSIS (FOR MEASURS WITH EXCLUSIONS --- gap in visits and medical visit frequency)**

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

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Not applicable.

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Not applicable.

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Not applicable.  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

**No risk adjustment or stratification**

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

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

**Other,** Click here to enter description

**2b4.1.1 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.**

Not applicable.

**2b4.2. 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**.

The Ryan White HIV/AIDS Program provides a comprehensive system of care that includes primary medical care and essential support services for people living with HIV who are uninsured or underinsured. The Program works with cities, states, and local community-based organizations to provide HIV care and treatment services to more than half a million people each year. The Program reaches approximately 52% of all people diagnosed with HIV in the United States.

As indicated in data presented earlier, the Ryan White HIV/AIDS Program is a public health, safety net program providing care to a high proportion of racial/ethnic minority, transgender, unstable housing, and low income people living with HIV. Many of people served by the Ryan White HIV/AIDS Program represent sociodemographics factors incorporate in risk adjusting models by many measures stewards.

As a result, the Ryan White HIV/AIDS Program does not adjust for risk in its performance measures. Rather, it is a fundamental aspect of the Ryan White HIV/AIDS Program to identify disparities and work to improve quality of care for subpopulations. Additionally, this measure is not used for pay-for-performance, bonuses, or penalties.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic 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*)

Not applicable.

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

Not applicable.

**2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS 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)**

Not applicable.

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

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.

Not applicable.

***If stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:** Not applicable.

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

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

**2b4.9. Results of Risk Stratification Analysis**: Not applicable.

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

Not applicable.

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

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

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

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

The chart-abstracted version of this measure has been in use since 2010. To examine meaningful differences in performance, we examined the distribution of the proportion of patients with viral suppression across providers, by year. Performance scores were broken into the bottom 10% and top 90% providers to better characterize the gaps that remain across providers. Moreover, performance scores were examined with respect the proportion of providers with least 80 percent of patients that were prescribed ART in a given year.

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | % patients with viral suppression across providers | | | | |  | providers with ≥80% patients prescribed ART | | |
| Year | Mean | SD | Median | 10th %ile | 90th %ile |  | N | n | % |
| 2010 | 65.9% | 27.5% | 76.5% | 17.8% | 91.2% |  | 846 | 353 | 41.7 |
| 2011 | 70.1% | 26.4% | 79.8% | 26.1% | 93.2% |  | 811 | 402 | 49.6 |
| 2012 | 73.4% | 25.4% | 83.8% | 31.7% | 94.7% |  | 816 | 471 | 57.7 |
| 2013 | 77.5% | 24.1% | 86.5% | 42.9% | 96.4% |  | 823 | 532 | 64.6 |
| 2014 | 78.0% | 28.0% | 90.0% | 29.6% | 98.3% |  | 813 | 565 | 69.5 |

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

The table above demonstrates meaningful variablility across providers, allowing for the identification of meaningful differences across sites. Specifically, the measure is able to detect providers with better or worse than median performance scores. In 2014, the bottom 10% of providers had ART prescription rates of 29.6% or lower; the top 90% of providers had rates of 98.3% or higher. These differences demonstrate the continued value of the measure in identifying sites based on poor performance relative to the top performers.

Provider-level performance differences observed in the table above also underscore improvements in the proportion of patients prescribed ART. In 2014, of 813 providers, 565 (69.5%) had prescribed ART for at least 80% of patients. Additionally, on average by provider, nearly 80% (78%) of patients were prescribed ART; however, given the large population that the RWHAP serves, even the poorest performing sites (e.g., bottom 10%) represent a substantial number of patients.

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

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

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

**Note***: This item is directed to measures that are risk-adjusted (with or without SDS 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 SDS 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.***

**2b6.1. 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*)

Not applicable

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

Not applicable

**2b6.3. What is your interpretation of the results in terms of 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*)

Not applicable

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

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

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

The HQMF standard specifies that if data are unknown or missing, they shall fail the criterion. This constraint embodies the notion that absence of evidence is evidence of absence, i.e. data not present in a structured field from which the measure draws will not be considered for measure calculation. In certain cases, missing data may have no impact on the measure outcome for a given patient. For example, a data element used in a series of OR statements will not impact the measure outcome if another data element in the OR statement is present and meets all other defined constraints.

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

The Bonnie synthetic patient bundle includes scenarios for missing data elements, which are a form of negative testing. All Bonnie synthetic patients with missing data performed according to the HQMF standard specification and as expected.

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

Please see response for question 2b7.1 above.