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

**Measure Number** (*if previously endorsed*)**:** 0027

**Measure Title**: Medical Assistance with Smoking and Tobacco Use Cessation

**Date of Submission**: 12/2/2016

**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: Patient-Reported Survey | other: Patient-Reported Survey |

**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*).

N/A

**1.3. What are the dates of the data used in testing**? Initial testing of measure score reliability was performed using HEDIS performance measurement 2010 data. For the 2016 update, we assessed measure score reliability using data from all health plans that submitted HEDIS data to NCQA for this measure and had a valid rate in 2015/2016, which used data submitted to NCQA in 2016. NCQA conducted cognitive testing of the questions in 2008.

**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: 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*)

2016 update: measure score reliability and construct validity testing: Measure score reliability was calculated for the three rates from the 59 commercial health plans, 159 Medicaid health plans, and 238 Medicare health plans that submitted data on this measure to HEDIS in 2016 (58 commercial plans submitted valid rates for the Discussing Cessation Medications and Discussing Cessation Strategies rates). Construct validity was calculated using the same commercial and Medicaid health plans (construct validity could not be calculated for the Medicare plans as they only submit one rate, Advising Smokers to Quit). Commercial and Medicaid plans are included in these analyses only if their denominator contains at least 100 individuals; Medicare plans must have a minimum of 30 individuals in their denominator to be included. The plans were geographically diverse and varied in size.

Systematic evaluation of face validity: This measure was tested for face validity with two panels of experts. Measurement Advisory Panels and subject matter workgroups provide the clinical and technical knowledge required to develop the measures. The Smoking Measurement Workgroup included nine experts in smoking and tobacco use and included representation from consumers, health plans, health care providers and policy makers. NCQA’s Committee on Performance Measurement (CPM) oversees the evolution of the HEDIS measurement set and includes representation from 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 is responsible for advising NCQA staff on the development and maintenance of performance measures. The CPM also meets with the NCQA Board of Directors to recommend measures for inclusion in HEDIS. 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. Additional HEDIS Expert Panels and the Technical Advisory Group provide invaluable assistance by identifying methodological issues and giving feedback on new and existing measures. See *Additional Information: Ad.1. Workgroup/Expert Panel Involved in Measure Development* for names and affiliations of expert panel members.

Initial testing of the CAHPS survey instrument:

There are two different and complementary approaches to assessing the reliability and validity of a questionnaire:

1. Cognitive testing, which bases its assessments on feedback from interviews with people who are asked to react to the survey questions.

2. Psychometric testing, which consists of analyses of data collected using the questionnaire.

**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*)

2016 update: measure score reliability and construct validity testing: In 2016, HEDIS measures covered 114.2 million commercial health plan beneficiaries, 47.0 million Medicaid beneficiaries, and 17.6 million Medicare beneficiaries. Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes number of health plans included in HEDIS data collection and the median eligible population for the measure across health plans.

Table 1. Denominator sizes for reliability and construct validity testing

|  |  |  |  |
| --- | --- | --- | --- |
| Product type | Rate | Number of plans | Median number of eligible patients per plan |
| Commercial | Advising Smokers to Quit | 59 | 132 |
| Discussing Cessation Medications | 58 | 132 |
| Discussing Cessation Strategies | 58 | 132 |
| Medicaid | Advising Smokers to Quit | 159 | 257 |
| Discussing Cessation Medications | 159 | 256 |
| Discussing Cessation Strategies | 159 | 256 |
| Medicare | Advising Smokers to Quit | 238 | 55 |

Patient sample for cognitive testing of survey questions: A total of 18 respondents were interviewed across two rounds of cognitive testing; age ranged from 26 to 69 years of age. Respondents were recruited for variation in race/ethnicity, level of smoking, and type of insurance—commercial, Medicare, and Medicaid.

**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 samples are described above.

**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).**

2016 update: Measure performance results are stratified by commercial, Medicaid and Medicare health plans.

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

**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*)

Method for initial measure score reliability testing: 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*)

2016 update: results for measure score reliability testing:

Table 2. Beta-binomial statistic for each measure rate

|  |  |  |
| --- | --- | --- |
|  | Commercial | Medicaid |
| Advising Smokers to Quit | 0.69 | 0.75 |
| Discussing Cessation Medications | 0.72 | 0.83 |
| Discussing Cessation Strategies | 0.77 | 0.77 |

Results for initial measure score reliability testing:

1. Commercial plans 2010:

1.a. Advising Smokers & Tobacco Users to Quit: 0.618960

1.b. Discussing Cessation Medications: 0.469075

1.c. Discussing Cessation Strategies: 0.700878

2. Medicaid 2010:

2.a. Advising Smokers & Tobacco Users to Quit: 0.605218

2.b. Discussing Cessation Medications: 0.851239

2.c. Discussing Cessation Strategies: 0.790183

3. Medicare 2010

3.a. Advising Smokers & Tobacco Users to Quit Only: 0.95

**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?*)

2016 update: interpretation of results for measure score reliability testing: Generally, a reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities. Beta binomial testing for this measure suggests that the three indicators within this measure have demonstrated good reliability.

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

**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)*

2016 update: method for assessing construct validity: We tested for construct validity by exploring whether the rates within this measure were correlated with each other, since they assess different aspects of tobacco use cessation. We hypothesized that organizations that perform well on one rate should perform well on the other rates. To test these correlations we used a Pearson correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variables 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 this measure, we specifically hypothesized:

1. The Advising Smokers to Quit rate will be positively correlated with the Discussing Cessation Medications rate.
2. The Advising Smokers to Quit rate will be positively correlated with the Discussing Cessation Strategies rate.
3. The Discussing Cessation Medications rate will be positively correlated with the Discussing Cessation Strategies rate.

To note, Medicare plans are not included in these analyses, because they only report the Advising Smokers to Quit rate.

Method for initial systematic assessment of face validity:

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.

Method for initial assessment of CAHPS survey instrument: Cognitive testing provides useful information about respondents’ comprehension of the questions, their ability to answer the questions, and the adequacy of the response choices. It also helps identify words that can be used to describe health care providers accurately and consistently across a range of consumers (e.g., commercially insured, Medicaid, fee-for-service, managed care, lower socioeconomic status (SES), middle SES, low literacy, higher literacy) and explores whether key words and concepts work equally well in both English and Spanish.

Field tests and psychometric analyses provide information about the items’ reliability and validity. Many existing questionnaires about health care have been tested primarily or exclusively using a psychometric approach, but the CAHPS team views the combination of cognitive and psychometric approaches as essential to producing the best possible survey instrument.

Additional information about cognitive testing of survey questions: The intent of the cognitive testing NQCA conducted in 2008 was to test four smoking cessation survey items from CAHPS 4.0 to address issues raised by health plans and the expert workgroup during the re-evaluation of the Medical Assistance with Smoking Cessation HEDIS measure. The recommended revisions to the questions addressed relevance with current guidelines (at the time), issues with response bias and clarity of survey language.

Cognitive testing was conducted in two rounds. In Round 1, all respondents were tested using the paper and pencil instrument (PAPI). Two versions of the protocol were tested in Round 2, a PAPI version and a telephone instrument version. The PAPI version was administered to the first four respondents and the telephone instrument was administered to the remaining five respondents.

For the PAPI testing, respondents were asked to read each item aloud and to provide a response. Probing was conducted concurrently, that is, administered after respondents answered each question. Interviewers administered the probes contained in the scripted protocol and followed up on other topics that emerged in testing. Respondents were given a $50 incentive to thank them for their time.

For the telephone instrument, the interviewer explained to the respondent that this is a survey that will be conducted by telephone. To simulate a telephone survey environment, the interviewer dialed in via phone to the respondent from another location. The interviewer read each question aloud to the respondent and noted the responses. After completing the survey by phone, the interviewer re-entered the room and administered the protocol retrospectively. For ease of reference, the respondent was able to look at each question and the answer they had provided.

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

2016 update: results from assessment of construct validity: The results indicate that the tobacco cessation rates are significantly correlated with each other in the direction that was hypothesized.

Results of Pearson correlation coefficient analyses (commercial health plans)

* Advising Smokers to Quit and Discussing Cessation Medications rates
  + Pearson correlation coefficient: 0.82
  + P-value: <.0001
* Advising Smokers to Quit and Discussing Cessation Strategies rates
  + Pearson correlation coefficient: 0.77
  + P-value: <.0001
* Discussing Cessation Medications and Discussing Cessation Strategies rates
  + Pearson correlation coefficient: 0.85
  + P-value: <.0001

Results of Pearson correlation coefficient analyses (Medicaid health plans)

* Advising Smokers to Quit and Discussing Cessation Medications rates
  + Pearson correlation coefficient: 0.74
  + P-value: <.0001
* Advising Smokers to Quit and Discussing Cessation Strategies rates
  + Pearson correlation coefficient: 0.68
  + P-value: <.0001
* Discussing Cessation Medications and Discussing Cessation Strategies rates
  + Pearson correlation coefficient: 0.84
  + P-value: <.0001

Results for initial systematic assessment of face validity:

Step 1: The Medical Assistance with Smoking Cessation was reevaluated in 2008. NCQA’s Performance Measurement Department and the Smoking Cessation Measurement Workgroup worked together to rename the measure to Medical Assistance with Smoking and Tobacco Use Cessation and review the cognitive testing results of the CAPHS survey.

Step 2: The proposed measure revisions and cognitive testing results were presented to the CPM in 2009. The CAHPS Survey has been deemed valid as a survey instrument. The CPM recommended to send the measure to public comment with a vote of 12 in favor and none opposed.

Step 3: The measure was released for Public Comment in spring 2009. We received and responded to comments on this measure. The CPM recommended moving this measure to first year data collection with a vote of 10 in favor and none opposed.

Step 4: The Medical Assistance with Smoking and Tobacco Use Cessation measure was introduced in HEDIS 2010. Organizations reported the measures in the first year and the results were analyzed for public reporting in the following two years. The Cardiovascular MAP assumed the responsibilities of the Smoking Cessation Measurement Workgroup reviewed the first year results and recommended public reporting. The CPM recommended moving this measure public reporting with a vote of 10 in favor 1 opposed, and 1 abstained.

Results from cognitive testing of survey questions: For the first topic, related to an individual’s current smoking status, respondents were asked to answer two questions: one that asked about cigarette use only, and one that asked about cigarette or tobacco use. Respondents were able to interpret the term “tobacco”, had no difficulty with the question, and preferred the question that used both smoking and tobacco to the question that used only smoking.

For the second topic, related to an individual’s experience with receiving advice to quit smoking by a doctor or other health provider, respondents were asked to answer two questions: one that used response options about the number of visits (eg, “1 visit”, “2 to 4 visits”, etc.) and one that used open quantifier response options (eg, “never”, “once”, “sometimes”, etc.). Respondents were generally more comfortable with the open quantifier response options, strongly preferring those response options to the options using number of visits. Respondents reported that the open quantifier response options were easier, and did not require remembering a concrete number of visits.

Similar to the second topic, for the third topic (related to an individual’s experience with discussing cessation medications) and fourth topic (related to an individual’s experience with discussing cessation methods other than medication), respondents were asked to answer two questions: one that used response options about the number of visits and one that used open quantifier response options. As with the second topic, respondents preferred the open quantifier response options. Respondents found it helpful that the questions included examples of cessation medications and cessation methods other than medication.

**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?*)

2016 update: interpretation of results from assessment of construct validity: Coefficients with absolute value of less than 0.3 are generally considered indicative of weak associations whereas absolute values of 0.3 or higher denote moderate to strong associations. The significance of a correlation coefficient is evaluated by testing the hypothesis that an observed coefficient calculated for the sample is different from zero. The resulting p-value indicates the probability of obtaining a difference at least as large as the one observed due to chance alone. We used a threshold of 0.0001 to evaluate the test results. P-values less than this this threshold imply that it is unlikely that a non-zero coefficient was observed due to chance alone. The results confirmed the hypotheses that the tobacco cessation rates are positively correlated with each other, suggesting they represent the same underlying quality construct of tobacco cessation care.

Interpretation of results from cognitive testing of survey questions: Respondents were able to answer the survey questions. They were able to interpret the addition of “tobacco” and they preferred open quantifier response options to response options that required them to recall the number of visits at which they discussed smoking or tobacco use cessation.

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

**2b3. EXCLUSIONS ANALYSIS**

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

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

**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*)

**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*)

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

**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.**

**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**.

**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*)

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

**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)**

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

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

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

**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*)

**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)*

2016 update: 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 .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 measured entities. However, the method can be used for comparison of any two measured entities.

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

**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*)

2016 update: The p-value for commercial and Medicaid plans was <0.001 for all three rates, except for the commercial and Medicare Advising Smokers to Quit rates (p-value <0.01).

Table 3. HEDIS 2016 variation in performance across health plans

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Product Line | Rate | Avg. EP | Avg. | SD | 10th | 25th | 50th | 75th | 90th | IQR | p-value |
| Commercial | Advising Smokers to Quit | 132 | 75% | 7% | 66% | 70% | 75% | 79% | 83% | 9% | 0.009 |
| Discussing Cessation Medications | 132 | 48% | 8% | 41% | 43% | 48% | 52% | 61% | 9% | <0.001 |
| Discussing Cessation Strategies | 131 | 44% | 9% | 34% | 38% | 42% | 50% | 58% | 12% | <0.001 |
| Medicaid | Advising Smokers to Quit | 257 | 76% | 6% | 68% | 73% | 77% | 79% | 82% | 6% | <0.001 |
| Discussing Cessation Medications | 256 | 48% | 8% | 37% | 43% | 48% | 54% | 58% | 11% | <0.001 |
| Discussing Cessation Strategies | 256 | 43% | 7% | 34% | 39% | 44% | 48% | 52% | 9% | <0.001 |
| Medicare | Advising Smokers to Quit | 55 | 86% | 6% | 78% | 82% | 86% | 90% | 93% | 8% | 0.006 |

EP: eligible population, the average denominator size across plans submitting to HEDIS

SD: standard deviation

IQR: interquartile range

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

Previous submission:

Commercial

ASTQ Rolling Average Rate

Data Element; 2010, 2009; 2008

N; 234; 14; 114

MEAN; 74.9; 79.5; 76.7

STDEV; 6.88; 5.99; 5.53

P10; 66.1; 73.8; 69.2

P25; 70.5; 73.9; 73

P50; 74.6; 79.8; 76.9

P75; 80.0; 84; 80.3

P90; 83.7; 87.7; 83.5

DSCM Rolling Average Rate

Data Element; 2010; 2009; 2008

N ; 231; 14; 115

MEAN; 50.5; 53.3; 54.4

STDEV; 7.8; 6.57; 7.29

P10; 40.3; 46.8; 45.2

P25; 45.3; 47.8; 48.7

P50; 50.3; 51.6; 53.6

P75; 55.5; 57.5; 61

P90; 62.0; 64.4; 63.6

DSCS Rolling Average Rate

Data Element; 2010; 2009; 2008

N; 228; 14; 115

MEAN; 42.8; 50; 49.7

STDEV; 8.7; 8.45; 7.46

P10; 33.0; 38.6; 39

P25; 35.9; 45.4; 44.6

P50; 41.2; 51; 49.7

P75; 47.8; 56.1; 55

P90; 55.1; 61.2; 59.8

Medicare

ASTQ Rolling Average Rate

Data Element; 2009

N; 295

MEAN; 78

STDEV; 7.91

STDERR; 0.46

MIN; 50

MAX; 100

P10; 67.7

P25; 73.3

P50; 78

P75; 82.9

P90; 87.8

Medicaid

ASTQ Rolling Average Rate

Data Element; 2010; 2009; 2008

N; 119; 99; 101

MEAN; 73.7; 74.3; 69.3

STDEV; 6.08; 5.3; 0.62

P10; 64.7; 67.1; 61.4

P25; 69.9; 70.8; 66.5

P50; 74.8; 74.9; 70.4

P75; 78.0; 77.7; 73.5

P90; 80.8; 80.8; 76.2

DSCM Rolling Average Rate

Data Element; 2010; 2009; 2008

N; 119; 99; 101

MEAN; 42.8; 43.4; 40.6

STDEV; 9.1; 9.38; 8.48

P10; 30.2; 29.4; 31.8

P25; 36.4; 37.2; 34.6

P50; 42.8; 43.4; 39.9

P75; 48.9; 51; 46.1

P90; 55.0; 56.6; 52.3

DSCS Rolling Average Rate

Data Element; 2010; 2009; 2008

N; 119; 98; 101

MEAN; 38.6; 38.8; 40.8

STDEV; 7.3; 7.78; 6.99

P10; 33.0; 28.4; 32.1

P25; 33.7; 34; 36.3

P50; 38.3; 38.3; 39.8

P75; 43.8; 44.4; 45.8

P90; 48.5; 50; 50.3

**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?*)

2016 update: The results above indicate that there is a six to 12 percent gap in performance between the 25th and 75th performing plans. For all product lines the difference between the 25th and 75th percentile is statistically significant. The largest gap in performance is for the commercial plans on the Discussing Cessation Strategies rate, which show a 12 percentage point gap between 25th and 75th percentile plans. The next largest gap is for Medicaid plans on the Discussing Cessation Medications, which show an 11 percentage point gap between 25th and 75th percentile plans. Additionally, the difference in performance between plans in the 10th and 90th percentiles varies from 12 to 27 points across the three rates and product lines. Overall, these results demonstrate that there are meaningful differences in performance for all three rates across commercial, Medicaid and Medicare product lines.

Previous submission: This information was not provided in the previous submission.

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

**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*)

**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*)

**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*)

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

**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*)

NA

**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*)  
NA

**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*)

NA