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

**Measure Title**: Oral Evaluation, Dental Services

**Date of Submission**: 2/10/2014

**Type of Measure:**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | ☐ Outcome (*including PRO-PM*) |
| ☐ Cost/resource | ☐ **XProcess** |
| ☐ 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). |

|  |
| --- |
| **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.  **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.    **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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) 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**.  **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.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

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

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

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

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

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

The testing datasets were consistent with the measure specifications for the target populations and reporting entities. This measure was specified for administrative enrollment and claims data for children with private or public insurance coverage. We used data from five sources and refer to “program” level information and “plan” level information. We included data for publicly insured children in the Texas Medicaid, Texas CHIP, Florida CHIP, and Florida Medicaid programs as well as national commercial data from Dental Service of Massachusetts, Inc. Florida and Texas represent two of the largest and most diverse states. The two states also represent the upper and lower bounds of dental utilization based on dental utilization data available from the Centers for Medicare and Medicaid Services. The five programs collectively represent different delivery system models. The Texas Medicaid data represented dental fee-for-service, and Texas CHIP data reflected a single dental managed care organization (MCO). The Florida CHIP data included data from two dental MCOs. The Florida Medicaid data include dental fee-for-service and prepaid dental data. The commercial data included members in indemnity and preferred provider organization (PPO) product lines.

**1.3. What are the dates of the data used in testing**? We used data from calendar years 2010 and 2011 for all programs except Florida Medicaid. Full-year data for 2011 were not available for Florida Medicaid.

**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 |
| **☐ X health plan** | **☐ X health plan** |
| **☐ X other: Program (e.g., Medicaid, CHIP)** | **☐ X other: Program (e.g., Medicaid, CHIP)** |

**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*)   
**Level of Analysis: Program, 5 Measured Entities**

1. Texas Medicaid
   1. Size: # Members 0-20 years, CY 2011: 3,554,247; # Members 0-20 years, CY 2010: 3,393,963

B. Location: Texas – Statewide

C. Delivery Type – FFS

2. Texas CHIP

1. Size: # Members 0-20 years, CY 2011: 842,454; # Members 0-20 years, CY 2010: 786,070

B. Location: Texas – Statewide

C. Delivery Type – Dental MCO (1 plan)

3. Florida CHIP

1. Size: # Members 0-20 years, CY 2011: 317,146; # Members 0-20 years, CY 2010: 315,975

B. Location: Florida – Statewide

C. Delivery Type – Dental MCO (2 plans)

4. Commercial

1. Size: # Members 0-20 years, CY 2011: 184,152; # Members 0-20 years, CY 2010: 189,968

B. Location: National

C. Delivery Type – Indemnity/FFS & PPO product lines

5. Florida Medicaid

1. Size: # Members 0-20 years, CY 2010: 2,068,670;
2. Location: Florida – Statewide

C. Delivery Type – FFS and Prepaid Dental

**Note:** At the time of testing, complete data were not available for Florida Medicaid for CY 2011.

**Level of Analysis: Plan, 2 Measured Entities**

The FL CHIP program had two separate dental plans that participate in the program in 2010 and 2011. Technically, we had three plans represented because the Texas CHIP program was served by a single dental plan so the program=plan in that case. For the purposes of testing plan comparisons within a program, we focus on the two plans in FL CHIP.

1. FL CHIP – Plan 1
   1. Size: # Members 0-20 years, CY 2011: 140,986; # Members 0-20 years, CY 2010: 77,255

B. Location: Florida – Statewide

C. Delivery Type – Dental MCO

1. FL CHIP – Plan 2
2. Size: # Members 0-20 years, CY 2011: 168,191; # Members 0-20 years, CY 2010: 116,388

B. Location: Florida – Statewide

C. Delivery Type – Dental MCO

**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*)

Note that there were only four programs in CY 2011 because Florida Medicaid did not have complete claims data available for CY 2011 at the time testing was conducted.

**Table 1.6A, Patient Characteristics, 0-20 Years Old, 2011**

****

**Table 1.6B, Patient Characteristics, 0-20 Years Old, 2010**

****

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

These data were used for all testing aspects except two:

A. Part of the face validity assessments involved expert consensus processes, including conducting an environmental scan of measure concepts and using the RAND-UCLA modified Delphi process to rate the importance, feasibility and validity. Please see section 2b2.2 for a complete description.

B. Data element validation using medical chart reviews did not include all programs. Due to the cost of these activities, chart reviews were conducted only for the Texas Medicaid and CHIP programs. Texas has the third largest Medicaid program and second largest CHIP in the U.S., both with significant diversity represented. In addition, the research team conducting the testing is the External Quality Review Organization for Texas and has years of experience conducting medical chart audits for the Texas Medicaid and CHIP programs for ongoing quality assurance purposes. Thus, an established infrastructure and expertise was in place to conduct chart reviews for these programs.

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

**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*)  
☐ **XCritical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)   
☐ **XPerformance 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*)

**Data Elements:**

* See section 2b2 for validity testing of data elements.
* Note: Unlike measures that rely on medical record data for which issues such as inter-rater reliability are likely to introduce measurement concerns or measures that rely on survey data for which issues such as internal consistency may be a concern, this measure relies on standard data fields commonly used in administrative data for a wide range of billing and reporting purposes.

**Measure Score – Threats to Measure Reliability**

An important component of assessing reliability is assessing, testing, and addressing threats to measure reliability.

**1. Evaluation of Clarity and Completeness of Measure Specifications**

For a measure to be reliable – to allow for meaningful comparisons across entities – the measure specifications must be unambiguous: the denominator criteria, numerator criteria, exclusions, and scoring need to be clearly specified. The initial measure specifications were developed by the Dental Quality Alliance (DQA). The Dental Quality Alliance includes 30 members, representing a broad range of stakeholders, including federal agencies involved with oral health services, dental professional associations, medical professional associations, dental and medical health insurance commercial plans, state Medicaid and CHIP programs, quality accrediting bodies, and the general public. The initial specifications were developed based on (1) evidence-based guidelines regarding the periodicity of oral evaluations, (2) an environmental scan that identified existing measure concepts and their limitations and (3) face validity assessments of the measure concept. These specifications were contained in the competitive Request for Proposals to conduct measure testing; a research team from the University of Florida was selected to conduct testing. The research team independently carefully evaluated whether the measure specifications identified all necessary data elements to calculate the numerators and denominators for each measure. In addition, the research team carefully reviewed the logic flow and made revision recommendations to improve the reliability of the resulting calculations. The DQA also solicited public comment on an Interim Report and posted the measurement specifications online for public comment. The research team worked with the DQA to evaluate and address all comments provided. Throughout the eight-month testing period, there were numerous reviews and revisions of the specifications conducted jointly by the research team and the DQA to ensure clear and detailed measure specifications.

**2. Sensitivity Testing of Measure Specifications**

Sensitivity testing included evaluating different measurement years (e.g., calendar year versus federal fiscal year). The measure score differences were less than one percentage point and were robust to the measurement year.

**3. Other Threats to Reliability - Sample Size**

Our measured entities include very large numbers of patients; therefore, small sample size is not a concern.

**2a2.3. For each level checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
See section 2b2 for validity testing of data elements.

**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?*)  
See section 2b2 for validity testing of data elements.

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

**2b2. VALIDITY TESTING**

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

☐ **Performance measure score**

☐ **Empirical validity testing**☐ **XSystematic 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 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)*  
We assessed (1) critical data element validity, (2) measure score validity, and (3) potential threats to validity.

**1. CRITICAL DATA ELEMENT VALIDITY**

Oral evaluation measures the percentage of children who received a comprehensive or periodic oral evaluation using procedure codes in administrative claims data to identify clinical oral evaluations. Thus, assessing the accuracy of procedure codes reported in the claims data is essential. The critical data elements for this measure include: (1) member ID (to link between claims and enrollment data), (2) date of birth, (3) monthly enrollment indicator, (4) date of service, and (5) Current Dental Terminology (CDT) codes. The first four items are core fields used in virtually all measures relying on administrative data and essential for any reporting or billing purposes. As such, it was determined that these fields have established reliability and validity. Thus, critical data element validity testing focused on assessing the accuracy of the dental procedure codes reported in the claims data as the data elements that contribute most to the measure score. To evaluate data element validity, we conducted reviews of dental records for the Texas Medicaid and CHIP programs. Validation of clinical codes in administrative claims data are most often conducted using manual abstraction from the patient’s full chart as the authoritative source. As described in detail below, we evaluated agreement between the claims data and dental charts by calculating the sensitivity, specificity, positive predictive value, and negative predictive value as well as the kappa statistic.

**A. Data Sources**

A random sample of encounters for members ages 3-18 years with at least one outpatient dental visit was selected for dental record reviews. The targeted number of records was 400. The expected response rate for returning records was 65%. Therefore, 600 records were requested. All outpatient dental records for members during an eight-month period were requested. Table 2b2.2-1 below summarizes the number of records requested and received. The number of eligible records received (414) exceeded the total targeted number of 400 records.

**Table 2b2.2-1 Dental Records Requested and Received**



**B. Record Review Methodology**

There were two components to the record reviews used to evaluate data element validity:

1. Encounter data validation (EDV) that provided an overall assessment of the accuracy of dental procedure codes found in the administrative claims data compared to dental records for the same dates of service.
2. Validation of oral evaluation procedure codes specifically.

The record reviews were conducted by two coders certified as registered health information technicians (RHITs). At weekly intervals during the record review process, the two RHITs randomly selected a sample of records to evaluate inter-rater reliability. A total of 100 records and 1,830 fields were reviewed by both individuals with 100% agreement.

**C. Encounter Data Validation – Overall Assessment**

For the first component of validation, encounter data validation, the research team followed standard Encounter Data Validation processes following External Quality Review protocols from CMS that it has used in ongoing quality assurance activities for the Texas Health and Human Services Commission. [Centers for Medicare and Medicaid Services, External Quality Review Encounter Data Validation Protocol (http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Quality-of-Care/Quality-of-Care-External-Quality-Review.html)]. The first three procedure codes were reviewed for each claim. A total of 1,135 procedure codes were reviewed. The RHITs were provided with a pre-populated data entry form with the codes from the claims data for the patient with the specified provider on a particular date of service. They evaluated whether the code in the claims data was supported by the dental record.

**D. Critical Data Element Validation – Oral Evaluation Procedures Codes**

**Data Extraction.** For the second component of validation, assessing whether oral evaluations are accurately captured by claims data, chart abstraction forms were developed by the research team to document evidence in the dental record that an oral evaluation had been performed. The chart abstraction forms and process were reviewed and approved by the DQA R&D Committee. Claims data were validated against dental records by comparing the dental records to the codes in the claims data for a randomly selected date of service. Prior to conducting the reviews, a sample of 30 records from prior encounter data validation activities was used to test the data abstraction tool and refinements were made accordingly. During the chart abstraction testing process, the RHITs met with the research team, which included two dentists (including a pediatric dentist), to review questions about interpreting the records. They then evaluated the 414 dental records using the data abstraction form. The results were recorded in an Access database. Specifically, the chart abstracting process involved identifying and recording whether there was any evidence of an oral evaluation being performed during the visit. The programming team extracted data from the administrative claims data for the same members and dates of service, recording the presence or absence of CDT codes for oral evaluations. The data files from the record review team and the programming team were merged into a single data file.

**Statistical Analysis.** To assess validity, we calculated sensitivity (accuracy of administrative data indicating a service was received when it is present in the chart), specificity (accuracy of administrative data indicating a service was not received when it is absent in the chart), positive predictive value (extent to which a procedure that is present in the administrative data is also present in the charts), and negative predictive value (extent to which a procedure that is absent from the administrative data is also absent in the chart). Positive and negative predictive values are influenced by sensitivity and specificity as well as the prevalence of the procedure. Thus, interpretation of “high” and “low” values is not straightforward. In addition, although charts are typically used as the authoritative source for validating claims data, some question whether charts always represent an “authoritative” source versus being better characterized as a “reference” standard. The kappa statistic has been recommended as “a more ‘neutral’ description of agreement between the 2 data sources . . . .” (Quan H, Parsons GA, Ghali WA, Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data, Med Care, 2004;42(8):801-809.) Thus, the kappa statistic also was used to compare the degree of agreement between the two data sources. A kappa statistic value of 0 reflects the amount of agreement that would be expected to be observed by chance. A kappa statistic value of 1 indicates perfect agreement. Guidance on interpreting the kappa statistic is: <0 (poor/less chance of agreement; 0.00-0.20 (slight agreement); 0.21-0.40 (fair agreement); 0.41-0.60 (moderate agreement); 0.61-0.80 (substantial agreement); 0.81-0.99 (almost perfect agreement). (Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. Jun 1977;33(2):363-374.)

**2. MEASURE SCORE - FACE VALIDITY**

Face validity of this measure was assessed at several stages during the measure development and testing processes.

**A. Face Validity Assessment – Measure Development**

Face validity was systematically assessed by recognized experts. The Dental Quality Alliance (DQA) was formed at the request of the Centers of Medicare and Medicaid Services (CMS) specifically for the purpose of bringing together recognized expertise in oral health to develop quality measures through consensus processes. As noted in the letter from Cindy Mann, JD, Director of the Center for Medicaid & CHIP Services within CMS: “The dearth of tested quality measures in oral health has been a concern to CMS and other payers of oral health services for quite some time.” (See Appendix)

During the measurement development process, the DQA Research and Development Committee, purposely comprised of individuals with recognized and appropriate expertise in oral health to lead quality measure development, undertook an environmental scan of existing pediatric oral health performance measures, which involved the following: (1) Literature Search, (2) Measure Solicitation, (3) Review of Measure Concepts, (4)Delphi Ratings of Measure Concepts, (5) Scan Results Analysis, (6) Gap Analysis, (7) Identification of Measures. A more detailed description of this process, the findings and the resulting measure concepts that were pursued is provided in reports published by the DQA. (Dental Quality Alliance. Pediatric Oral Health Quality and Performance Measures: Environmental Scan. 2012; Dental Quality Alliance. Pediatric Oral Health Quality & Performance Measure Concept Set: Achieving Standardization & Alignment. 2012. Both reports available at: http://ada.org/7503.aspx. )

**(1) Literature Search.** The Committee began its work by identifying existing performance and quality measure concepts (description, numerator, and denominator) on pediatric populations defined as children younger than 21 years. Staff conducted a comprehensive online search for publicly available measure concepts. This search was conducted initially in August – September 2011 and then updated on February 8, 2012. The following searches were conducted: (1) PubMed Search. Staff used two specific search strategies to search Medline. Search 1: (performance OR process OR outcome OR quality) AND measure AND (oral or dental) AND (children OR child OR pediatric OR paediatric) – 1121 citations. Search 2 - "Quality Indicators, Health Care"[Mesh] AND (dental OR oral) - 150 citations. Staff included five articles based on title and abstract review of these citations. Measure concepts presented within these articles were included in the list of concepts for R&D Committee review. (2) Web Search. Staff then performed an internet search with keywords similar to the ones used for the PubMed search. (3) Search of relevant organization websites. Staff began this search through the links provided within the National Library of Medicine database of relevant organizations (<http://www.nlm.nih.gov/hsrinfo/quality.html#760>). Example of organizations involved in quality measurement include the National Quality Measures Clearinghouse (NQMC), National Quality Forum (NQF), and Maternal and Child Health Bureau (MCHB).

**(2) Solicitation of Measures.** In addition, the R&D Committee contacted staff at the Agency for Healthcare Research and Quality (AHRQ) in August 2011 to obtain the measures collected by the Subcommittee on Children’s Healthcare Quality for Medicaid and CHIP programs (SNAC). The Committee solicited measures from other entities, such as the DentaQuest Institute, involved in measure development activities.

**(3) Review of Measure Concepts.** Using inclusion/exclusion criteria, the R&D Committee reviewed the measure concepts and identified the measures that would be reviewed and rated in greater depth.

**(4) Delphi Ratings.** The RAND-UCLA modified Delphi approach was used to rate the remaining measure concepts, applying the criteria and scoring system for importance, validity, and feasibility consistent with the process that was used by the SNAC. There were two rounds of Delphi ratings to identify a starter set of pediatric oral health performance measures. [Brook RH. The RAND/UCLA appropriateness method. In: McCormick KA, Moore SR, Siegel R, United States. Agency for Health Care Policy and Research. Office of the Forum for Quality and Effectiveness in Health Care., editors. Clinical practice guideline development : methodology perspectives.]

**(5) Scan Results.** There were a total of 112 measure concepts identified through the environmental scan: 59 met the inclusion criteria for being processed through the Delphi rating process and 53 did not. Among the 59 measures that were evaluated through the Delphi rating process, 38 were deemed “low-scoring measure concepts” and 21 were deemed “high-scoring measure concepts.”

**(6) Gap Analysis.** The R&D Committee then identified the gaps in existing measures, including both gaps in terms of the care domains addressed (e.g., use of services, prevention, care continuity) as well as gaps based on good measurement practices (e.g., standardized measurement methodology, evidence-based, etc.). Although the Committee did identify content areas that were not addressed, a key finding was the lack of standardized, clearly-specified, validated measures.

**(7) Identification of Measures.** The findings were used to identify a starter set of measures that would achieve the following objectives: (a) uniformly assess the quality of care for comparison of results across private/public sectors and across state/community and national levels; (b) inform performance improvement projects longitudinally and monitor improvements in care; (c) identify variations in care, and (d) develop benchmarks for comparison.

**B. Face Validity Assessment – Measure Testing**

The research team and the DQA R&D Committee continued to assess face validity throughout the testing process. Face validity also was gauged through feedback solicited through public comment periods. In March 2013, an Interim Report describing the measures, testing process, and preliminary results was sent to a broad range of stakeholders, including representatives of federal agencies, dental professionals/professional associations, state Medicaid and CHIP programs, community health centers, and pediatric medical professionals/professional associations. Each comment received was carefully reviewed and addressed by the research team and DQA, which entailed additional sensitivity testing and refinement of the measure specifications. Draft measure specifications were subsequently posted on the DQA’s website in a public area and public comment was invited. National presentations, including presentations at the National Oral Health Conference, were made by the research team and DQA in the spring and summer of 2013, which included reference to the website containing the measure specifications and invitations to provide feedback. All comments received were reviewed and addressed by the research team and DQA, including additional sensitivity testing and refinement of the measure specifications.

The final face validity assessment was conducted at the July 2013 Dental Alliance Quality meeting at which the full membership, representing a broad range of stakeholders. A detailed presentation of the testing results was provided. The membership then participated in an open consensus process with observed unanimous agreement that the calculated measure scores can be used to evaluate quality of care.

Sample Presentations

Aravamudhan K. Dental Quality Alliance Measures. Presentation at 2013 National Oral Health Conference Pre-Conference Workshop on Objectives, Indicators, Measures and Metrics. 2013.

Herndon JB. DQA Pediatric Oral Health Performance Measure Set: Overview of Measures and Validation Process. Presentation at 2013 National Oral Health Conference Pre-Conference Workshop on Objectives, Indicators, Measures and Metrics. 2013.

Herndon JB. DQA Pediatric Oral Health Performance Measure Set: Overview of Measures and Validation Process. Presentation at 2013 Texas Medicaid and CHIP Managed Care Quality Forum. 2013.

**3. ADDITIONAL VALIDITY TESTING - DENOMINATOR ENROLLMENT CRITERIA**

To finalize the denominator definition, several different enrollment criteria were tested: (1) enrolled at least one month, (2) enrolled at least three months, (3) enrolled at least 6 months, (4) enrolled the entire year (12 months), allowing a single one-month gap, and (5) average period of enrollment/person-time equivalent (weighting members in denominator by enrollment length). These were evaluated through the face validity consensus processes.

The first definition was ruled out because of concern that one month is an insufficient period of time to expect children to seek, schedule, and obtain a dental visit. The last definition was ruled out on the basis of usability as it was considered to be less readily interpretable by a wide range of stakeholders. Table 2a2.2-2 summarizes the percentage of members enrolled in the program during the reporting year who were eligible under each of the different enrollment intervals. Table 2a2.2-3 summarizes the performance scores that were calculated using each of the enrollment criteria longer than one month. Based on these data, a consensus was reached to adopt a six-month continuous enrollment requirement to balance sufficient enrollment duration that allows children adequate time to access care (seek, schedule and obtain a dental visit) with the number of children who drop out of the denominator due to stricter enrollment requirements.

**Table 2b2.2-2. Percentage of All Enrolled Members Included in Different Denominator Definitions**



**Table 2b2.2-2. Performance Rates for Different Denominator Definitions**



**4. ADDITIONAL VALIDITY TESTING - CONVERGENT VALIDITY**

We also evaluated the extent to which the measure score demonstrated convergent validity (degree to which the measure score is similar to other measures of the same construct) by using data from the Centers for Medicare and Medicaid Services (CMS) Form 416 reports on EPSDT eligible children enrolled in Medicaid for at least 90 days who received “diagnostic dental services,” which includes all clinical oral evaluations (a broader set of oral evaluations than is included in the proposed Oral Evaluation measure). To address the differences in enrollment requirements (CMS requires 90 days and the proposed measure requires 6 months), we calculated the rates for the proposed measure using a 3-month enrollment criterion in order to compare the rates for the proposed measure to CMS-416 data for the Florida and Texas Medicaid programs. We used the CMS-416 data in to calculate the percentage of EPSDT eligible children enrolled at least 90 days who received “diagnostic dental services.”

**5. ADDITIONAL VALIDITY EVALUATION – ASSESSMENT OF THREATS TO VALIDITY**

**A. Exclusions**

As described in 2b3. of this form, there are no exclusions for this measure.

**B. Risk Adjustment**

Risk adjustment is not applicable for this process measure.

**C. Missing Data**

As described in measure evaluation criteria 3c1, this measure relies on standard data elements in claims data that are already collected and widely used for a range of reporting and billing purposes with very low rates of missing or invalid data (which we empirically assessed and reported in 3c1).

**D. Multiple Sets of Specifications**

This does not apply to the proposed measure.

**E. Ability to Identify Statistically Significant and Meaningful Differences in Performance**

As described in 2b5 of this form, this measure is able to identify statistically significant and meaningful differences in performance. We also demonstrate with empirical data and statistical testing the ability of this measure to detect disparities in 1b4 (Importance).

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

**1. CRITICAL DATA ELEMENT VALIDITY**

**A. Encounter Data Validation – Overall Assessment**

Encounter data validation of 1,135 procedure codes in the claims data against dental charts found agreement for 94% of the procedure codes (Table 2b2.3-1). Only 4.2% of procedure codes reported in the administrative data were not supported by evidence in the dental record. For 1.8% of the records reviewed, the documentation was insufficient to determine whether the service indicated by the procedure code had been rendered or not.

**Table 2b2.3-1 Agreement between Records and Administrative Data for Procedures**



**B. Critical Data Element Validation – Dental Service Procedure Codes for Oral Evaluations**

To assess whether oral evaluations performed are accurately captured by claims data, the 414 records, representing 631 dates of service, were reviewed. Table 2b2.3-2 below summarizes the agreement between the dental records and administrative data. Agreement (concordance) between the dental records and administrative claims data was 86.6%. Sensitivity was 85.1% and specificity was 92.5%. The positive predictive values was 97.9%, and the negative predictive value was 59.7%. As noted above, the kappa statistic provides a more neutral description of agreement and extends a comparison of simple agreement by taking into account agreement occurring by chance, thereby providing a more rigorous and conservative measure of agreement between the two data sources. The kappa statistic value was 0.642, indicating “substantial” agreement. Collectively, these findings indicate strong concordance with a greater likelihood of false negatives than false positives. Evaluating dental records for documented evidence oral evaluations was more challenging than identifying whether other specific procedures were performed, such as topical fluoride application or restorative procedures, because oral evaluations encompass a set of services and there is greater variability in charting practices related to documenting oral evaluations. The RHITs erred on the side of being over-inclusive in recording evidence of an oral evaluation, which may have contributed to the finding of a greater likelihood of false positives.

**Table 2b2.3-2 Agreement between Record and Administrative Data for Specific Care Domains**



We compared our findings to those in the peer-reviewed literature. A study was conducted in 2004 that used data from 3,751 patient visits in 120 dental practices participating in the Ohio Practice-Based Research Network to examine the concordance of chart and billing data with direct observation of dental procedures. They evaluated “oral examinations,” which were broadly defined. For oral examinations, they found lower sensitivity (42%), similar specificity (96%), and a lower kappa value (0.44). They noted, however, that the categories in the form they used to identify oral examinations through observation were general in nature and “included any activity that was used to determine the oral health or status of a patient from simple mouth mirror examinations to Diagnodent evaluation.” (Demko CA, Victoroff KZ, Wotman S. 2008. “Concordance of chart and billing data with direct observation in dental practice” Community Dent Oral Epidemiol. 36(5):466-74.)

**2. FACE VALIDITY**

**Oral Evaluation,** and specifically a comprehensive or periodic oral evaluation, was identified through the Delphi rating process as a high-scoring measure concept with a mean importance score of 8, mean feasibility score of 8, and mean validity score of 8, all out of a 9-point scale. [Rating of 1-3: not scientifically sound and invalid; 4-6 – uncertain scientific soundness and uncertain validity; 7-9 – scientifically sound and valid.] Median score ratings were equal to the mean ratings. Thus, the measure has face validity. However, gaps were identified with existing measures, including defining “diagnostic services” or “examination” too broadly, lack of clear specifications, and lack of standardization.

**3. MEASURE SCORE - CONVERGENT VALIDITY**

Measure score validity was further assessed by comparisons to the CMS EPSDT data for the Florida and Texas Medicaid programs, using the data in the Form 416 reports to calculate the percentage of EPSDT eligible children enrolled at least 90 days who received “diagnostic dental services,” which includes all clinical oral evaluations. (The CMS numerator includes periodic and comprehensive oral evaluations, but also problem-focused oral evaluations.) The rates calculated for the proposed Oral Evaluation measure using the test data (and 3-month instead of 6-month enrollment criteria) and those calculated using the CMS-416 Form data resulted in rates that were within 2 percentage points for the measure overall and for most of the age stratifications for both states (Table 2b2.3-3). Although the enrollment duration used for this comparison is different than that specified for the measures, our comparison of rates by enrollment duration demonstrated fairly consistent increases in the rates across the programs with an increase in the enrollment criterion from 3 months to 6 months. Therefore, we believe the similarities in the rates for the 3-month enrollment criteria provide evidence of convergent validity.

**Table 2b2.3-3 Comparison of DQA Oral Evaluation Measure Score to Similar Domain Calculated using CMS Form 416 EPSDT Data**



**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?*)  
Both face validity and convergent validity of the measure scores were established. For the critical data elements, there was strong overall concordance between the administrative claims data and dental records and “substantial” agreement based on the more conservative Kappa statistic. Collectively, these findings lead us to conclude that the measure score represents a valid measure of oral evaluations.

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

**2b3. EXCLUSIONS ANALYSIS**

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

The only exclusions were those that are standard exclusions in any measure reporting: children who do not qualify for dental benefits under their coverage were not included because this measure is intended only for children with dental coverage. For example, individuals 0-20 years with Medicaid coverage for emergency services only or for pregnancy-related services that do not provide dental coverage were not included.

**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)***.***

Not applicable.

**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

Not applicable.

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.   
Not applicable.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient 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 and not related to disparities*)

**2b4.4. What were the statistical results of the analyses used to select risk factors?**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*)  
Not applicable.

*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*)**:**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*)

Not applicable.

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

**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)*

This is a new measure. As noted in 1b, there were variations in the measure scores across the five programs included in the testing. For convenience we have included the performance score data from 1b below. In addition to providing the 95% confidence intervals for each score, we used chi-square tests to analyze whether there were statistically significant differences between (1) the 4 programs with performance data for 2011, (2) the 5 programs with performance data for 2010, (3) the two dental MCOs in FL CHIP in CY 2010 and (4) the two dental MCOs in FL CHIP in CY 2011. Because the measure score is the proportion of children who received a service, the dichotomous outcome of had/did not have a service can be used to conduct chi-square significance testing in order to evaluate whether there are statistically significant differences in the measure scores between programs and between plans.

**Table 1b.2. Performance Scores**

Program, Year, Measure Score as % (Measure Score, SD, Lower 95% CI, Upper 95% CI)



**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*)

For both years, statistically significant differences were detected in the measure scores between programs and between plans (Table 2b5.2).

**Table 2b5.2. Chi-Square Test of Differences in Measure Scores**



**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?*)  
Statistically significant differences between measured entities were detected at both the program and plan reporting levels. We believe this is consistent with evidence reported elsewhere in this application documenting a performance gap and disparities in performance. Thus, this measure informs performance improvement efforts by allowing plans and programs to identify and monitor performance gaps both at any given point in time and over time.

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

**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 criterion is directed 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).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
 Not applicable.

**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 demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
Not applicable.